DOCUMENT RESUME

ED 346 133 TM 018 430

AUTHOR Baker, Eva L.
TITLE The Role of Domain Specifications in Improving the Technical Quality of Performance Assessments. Project 2.2: Alternative Approaches to Measuring Liberal Arts Subjects: History, Geography, and Writing.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Feb 92
CONTRACT R117G10027
NOTE 145p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Educational Assessment; Essay Tests; Grade 11; High Schools; *High School Students; Objective Tests; Rating Scales; *Scoring; Sex Differences; *Test Construction; Test Use; Test Validity; *United States History
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

Work on the development of history performance assessments is described. So far, six complete sets of assessments in United States history have been developed as part of this project. Students are first assessed on their historical knowledge of the period through a short-answer test. They are then asked to write an essay explaining the positions of authors of the primary source texts provided. A series of studies conducted to determine how scoring rubrics should be developed found the best strategy to involve looking at differences between expert and novice performance. The use of task specifications controlled some of the variation commonly associated with performance tasks. Results from the research to date include: (1) development of a valid scoring scheme; (2) development of rater training procedures; (3) a task structure that reduces score variability; (4) distinguishing between assessment purposes and the utility of overall score and subscores; (5) detecting gender differences in scores; (6) supporting the validity of the measures through grade point averages and a scale ensuring effort through studies with 68 11th graders; and (7) addressing validity criteria with the same data. There are 27 tables and 4 graphs of study data, and a 40-item list of references. Five appendices provide samples of the content assessments and information about validity and reliability results. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346133

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – February 1992

Project 2.2: Alternative Approaches to
Measuring Liberal Arts Subjects:
History, Geography, and Writing

The Role of Domain Specifications in Improving the Technical Quality of Performance Assessment

Eva L. Baker, Project Director

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM018430

BEST COPY AVAILABLE

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

THE ROLE OF DOMAIN SPECIFICATIONS IN IMPROVING THE TECHNICAL QUALITY OF PERFORMANCE ASSESSMENT

Eva L. Baker

Executive Summary

We have conducted research on the design of history performance tasks since 1988. Focusing on both explanation and knowledge representation skills, the project has attempted to determine how secondary students' deep understanding of history subject matter can be validly assessed. The performance tasks have evolved to require students to engage in a sequence of assessed steps which require a minimum of one-and-a-half hours a topic. First, students are assessed on their relevant background knowledge of the particular historical period. The measure consists of a 20-item, short-answer test with questions to measure principles and specific events directly relevant or less relevant to the historical issue under assessment. Next students are provided with opposing viewpoints in primary source text materials, typically letters or speeches of historical figures. Finally, students are asked, in a highly contextualized set of directions, to write an essay that explains the positions of the authors of the texts, and to draw upon their own background knowledge for explanation. In some studies, students have been given optional resources to read, or have been required to prepare HyperCard or concept map representations of the key knowledge, principles and relationships in the text materials (Baker, Niemi, Novak, & Herl, in press).

A series of studies was conducted to determine how scoring rubrics should be developed, and the best strategy relied on looking at differences between expert and novice performance (Baker, Freeman, & Clayton, 1991). The essay scoring rubric consists of six dimensions: a General Content Quality scale, focused on the overall quality of the content understanding, and five subscales—Prior Knowledge (the facts, information, and events outside the provided texts used to elaborate positions); Principles (the number and depth of description of principles); Text (the use of information from the text for elaboration); Misconceptions (the number and scope of misunderstandings in interpretation of the text and historical period); Argumentation (the quality of the argument, its logic and integration of elements). These dimensions are

scored on a 1-5 scale. History experts and high school teachers have been involved throughout the study as co-designers, reviewers, and raters of the assessment.

So far, six complete sets of history assessments have been developed: two on the Revolutionary Period; one on the Civil War; two on 20th century immigration; and one on the Depression Period. These tasks connect to the California *History-Social Science Framework* (California State Department of Education, 1988). Replications in the areas of science (Baker, Niemi, Novak, & Herl, in press) and economics, (Baker, 1991) have been conducted to assess the utility of the scoring rubric for explanation tasks in other content areas.

What Have We Learned From Our Research on History Performance Assessment?

1. We have developed a valid scoring scheme for assessing deep understanding. It supports the theoretical model under which it was developed in this replication. It is generalizable across history topics. It is ready for dissemination.

2. We have developed rater training procedures that produce reliable and valid scoring of student tasks in a limited period. The scoring rubric makes strong cognitive demands of the raters. The training procedure is ready for dissemination.

3. We have a task structure that reduces score variability such that fewer topics can be used to derive reliable scores for individual students. This technique is more efficient than found in most comparable studies. These relationships are all the more startling because of the lack of preparation and motivation among our students.

4. We have distinguished between assessment purposes and the utility of overall score and subscores.

5. We have found gender differences in this small sample, favoring females.

6. We have found supportive data for the validity of our measures in grade point average (GPA) and a scale measuring student effort.

7. We have systematically addressed validity criteria (Linn, Baker, & Dunbar, 1991) in our research studies: the criteria addressed include fairness, generalizability, cognitive complexity, content quality, reliability, cost, and efficiency. We are in the process of conducting studies of transfer and designing research to assess the meaningfulness of tasks to students.

THE ROLE OF DOMAIN SPECIFICATIONS IN IMPROVING THE TECHNICAL QUALITY OF PERFORMANCE ASSESSMENT

Eva L. Baker

Introduction

With little experience as a guide, in the world of educational testing performance assessment is speedily becoming the rhetorically preferred option to multiple choice and other traditional forms of achievement testing. Assessing students' performance of complex, challenging tasks is logically compelling. Teaching energy presumably becomes focused directly on student learning activities, the coin of the instructional realm, rather than allocated mainly to imparting information, following the textbook, or preparing for decontextualized tests. Support for new performance assessments also comes from research in the psychology of learning (Baker, O'Neil, & Linn, 1991; Resnick & Resnick, 1992; Shepard, 1991), which argues that the assessments can model attributes of constructivist and motivational learning theories.

These positions stimulate the proliferation of efforts to generate such new assessments. States and local school districts have begun to develop performance assessments in virtually all subject areas, in multidisciplinary contexts, and in applied settings such as workforce readiness (Aschbacher, 1991; Baker, 1990). Consortia through the Council of Chief State School Officers and the New Standards Project have cooperated to support test development. States such as California, New York, Michigan, Arizona, Connecticut, and Vermont have taken the lead, but many others are rapidly closing on them. Powerful assessment models emphasizing student focus and engagement have been demonstrated through the work of the Coalition of Essential Schools (Sizer, 1992), and Project Zero (Wolf, 1989). Commercial publishers are embarking on similar ventures.

The breadth and intensity of these efforts have been sustained by visible reports of national policymakers (*America 2000*, U.S. Department of Education, 1991; *What Work Requires of Schools*, U.S. Department of Labor,

1991; *Raising Standards for American Education*, National Council of Education Standards and Testing, 1992). They argue that these new forms of assessments are critical to the improvement of education services, the motivation of students to learn, and even to the restoration of U.S. economic leadership.

Although there is uniform agreement that performance assessments should focus on the improvement of learning, there are important differences in how we should proceed in the accountability area. Some argue that unless individual student progress is measured and reported to parents and policymakers, assessments will not matter. Others believe that it is desirable to expand the use of assessments for use in grade-to-grade promotion, college admissions, and job selection decisions. In a nutshell, many see the utility of these performance assessments increased as they emphasize personal and system accountability. Not everyone is convinced that unproven and largely unexplored measurement techniques should be made into national policy. In particular, the reversion of the military from their use of some forms of performance assessments to paper-and-pencil tests, and the recent problems in fielding a national performance assessment system in the schools of the United Kingdom (D. Nuttall, personal communication, 1992) have heightened worries about assessment utility.

So faint that it is almost missing from the rhetoric surrounding performance assessment is discussion of the technical quality of these assessments. In part, such concerns are difficult to put forth, since there has been relatively little technical research on performance assessments. CRESST, the National Center for Research on Evaluation, Standards, and Student Testing, in its proposal (Baker & Linn, 1990) articulated a set of criteria against which the validity of new performance assessments should be judged. Modified successively (Baker, 1990; Baker, O'Neil, & Linn, 1991; Linn, Baker, & Dunbar, 1992), these criteria set goals for the development of new and application of existing methodologies to the quality of these measures. These criteria include internal attributes of assessments (their cognitive demands, content adequacy, motivational attributes) that can be judged by experts, and external criteria (such as validity for various purposes, reliability, fairness, and consequences) that require data collection and analysis. In addition, practical criteria, such as cost and feasibility, are also seen as critical.

A key focus for research in the next few years must be the relationship of design strategies, that is, how we make such performance tests, and the technical quality (validity, content adequacy, etc.) of resulting assessments. This report describes the relationship of performance assessment development processes and resulting quality of measures.

The research below was conducted to answer the following question:

> How does the use of controlled design of performance assessments relate to the subsequent technical quality of such assessments?

In particular, the following questions were posed:

1. How generalizable are these assessments?
2. How effectively can we control topic and rater variability?
3. How do technical analyses support the construction of the scoring rubric?
4. If these assessments were to be used to estimate a domain or universe of performance, how many separate assessments would need to be given?
5. How many raters would be required to achieve reasonable reliability?
6. What is the relationship of measured student performance and other indices of student motivation and achievement?
7. How fair are these assessments?

Rationale for a Performance Assessment Design Process

Many performance assessments are developed with minimal design constraints, for no clearly acknowledged technology exists for performance task design. Developers seem to focus on a few limits when they create new assessments. One set of constraints concerns logistical issues, such as assessment time and availability of materials. Another emphasis is the surface characteristic of the task, that it exhibits motivational or "authentic" attributes of the assessment. Teachers and other developers want these assessments to capture the imagination of students and teachers, to be intrinsically motivating, and if possible to be relevant to real-world demands and expectations. Far less attention has been paid to design constraints

focused on increasing the technical quality and the economic feasibility of the resulting assessments.

This research assumes that a desired goal of performance assessment is the generation of "comparable" tasks for estimating student achievement. Comparability has been identified as a key dimension in the design of performance assessments in a national policy context (National Council on Education Standards and Testing, 1992) and researchers have defined comparability in a number of ways (Baker, O'Neil, & Linn, 1991; Linn, 1991; Linn, Kiplinger, Chapman, & LeMahieu, in press). Standards for comparability depend upon how the assessment will be used. Task comparability is desirable if one's intended use of results is instructional improvement, since ideally teachers would wish to select and employ similar tasks in instruction to prepare students for criterion performance settings.

In high stakes contexts, where assessments are used to make decisions about students' access to future education or other life chances, many believe task comparability is essential. Chances for individual success must not depend upon the particular task or topic of any performance assessment. Tasks must be relatively interchangeable, in other words, with low person-by-task interactions. To establish general competency for an individual student, the findings from one performance task should apply to a larger universe of similar tasks the student might be asked to perform. The problem is obviously tied up with fairness as well, for we demand that chances of success for different students be independent of the particular performance task they were presented.

Although comparability in certain narrow cases can be achieved by scaling and statistical equating (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), there is little evidence to suggest that these techniques are appropriate for tasks with widely varying requirements and technical characteristics. The research reported herein sought to produce comparability by designing it at the outset rather than adjusting findings *post hoc*. Specifications to control the cognitive demands of the task, the structure of the assessments, and the generation and application of scoring rubrics were thought to produce performance that showed less variability from topic to topic

than tasks created with fewer design constraints. In our attempt, we tried to control both rater and score reliability.

The specification strategy we adopted is based on the use of task structures or domain specifications. The major sources for this assessment design strategy were Baker and Herman (1983), Millman and Greene (1989), Popham (1987), and Hively, Patterson, and Page (1968), all of whom urge, more or less, the creation of specifications of a domain of tasks from which individual assessments can be drawn. Baker and Herman argue for a common structure or task syntax for the assessments into which particular topics would be substituted. In the present research, task specifications include design of the prior knowledge test, selection of text materials, the essay prompt design to elicit explanation performance, and the use of a common scoring system. In summary, the specifications target a cognitive task domain, formats for the presentation of materials, scoring schemes, and procedures for training raters. Each specific task has a highly contextualized prompt and specific content relevant to a given historical period.

Procedures

Sixty-eight students in two 11th-grade history classes in a California school district were the subjects of the study. Students were predominantly middle-class whites, although approximately 28% were middle-class Asians. Teachers agreed to participate to gain an introduction to the ideas of performance assessment. Teachers were asked to administer the performance assessments and additional research instruments. As an incentive, an invitational training session at a university campus was scheduled including these and other high school history teachers.

The data were collected over a one-month period. First, the *Test Anxiety Inventory* (Spielberger, 1980) was given to determine the trait test anxiety for the subjects; a student information form asking about their coursework patterns was also given. Three performance task administrations, lasting about two hours each, occurred one week apart, involving three different topics. All students, in the first week, were administered the performance task relating to the Civil War (CW) (see Appendix A). For the second and third sessions, the Chinese Immigration (CI) (see Appendix B) and General

Immigration (GI) (see Appendix C) tasks were administered. These tasks were administered in a counterbalanced design, with administration order randomly assigned within classroom. This counterbalanced design was necessary to determine whether there were specific effects of moving from specific knowledge to general knowledge or vice versa in the immigration area. Each student ultimately completed the Civil War, the Chinese Immigration, and the General Immigration task.

Each assessment session was administered by the classroom teacher with observation by research staff. Teachers read test instructions and provided minimal logistical assistance to students. No subject matter or strategic help was provided to students during the testing periods. In each two-hour session, the following sequence of assessment materials was administered.

Table 1

Description of Student Assessment Tasks in Sequence

Measure	Format	Time (Minutes)
1. Prior Knowledge	20 short answer items	15
2. State Anxiety-1[a] (Worry)	5 item Likert	2
3. Primary Source Material	Two texts	30
4. Literal Comprehension	12-14 multiple-choice items	10
5. State Anxiety-2	5 item Likert	2
6. Explanation task	Essay	50
7. State Anxiety-3	5 item Likert	2
8. Thinking[b]	28 item Likert	5
9. Debriefing	18 items	5

[a] Morris, Davis, & Hutchings, 1981.

[b] O'Neil, Baker, Jacoby, Ni, & Wittrock, 1990.

Data were collected for 69* students on the Civil War task, 68 on the Chinese Immigration task, and 68 on the General Immigration task.

Scoring Task Performance

Prior knowledge measures were scored on a 0-4 scale by one rater. Previous experience (Baker, Clayton, Aschbacher, Chang, & Ni, 1990)

* Original number tested; one subject omitted due to incomplete data.

demonstrated that interrater reliabilities were in the .98 range and justified the cost savings associated with only one rater.

Essay performance was rated using the History Explanation Rubric described in previous literature (Baker, Aschbacher, Niemi, Chang, Weinstock, & Herl, 1991; Baker, Clayton, Aschbacher, Chang, & Ni, 1990, Baker, Freeman, & Clayton, 1991) with a modification. In this trial, we added a new, experimental subscale, "Argumentation," to measure the logical progression and integration of the principles, text, and prior knowledge categories. This is a modification of an earlier dimension, "Interrelationships," used in previous trials. Four raters were trained to use the scoring rubric (see Appendix D), using materials developed and validated for training raters (Baker, Niemi, Aschbacher, Ni, & Yamaguchi, 1991).

All of the raters were secondary history teachers. Training involved familiarization with the scoring rubric and scoring of prescored papers to illustrate the desired ranges for rubric score points. Training took approximately three hours and the criterion/reliability segment required about 45 additional minutes. Each rater scored the three tasks for all students. Papers on each topic were rated together. During the scoring period, three criterion checks were conducted by examining scores for previously-scored papers that had been inserted into each rater's set, to assure that raters were maintaining appropriate score point definitions. Raters also scored a randomly selected subsample of 17 papers twice to permit analysis of intrarater reliability.

Analyses and Results

Means and standard deviations were computed for each student on the preliminary elements of each performance topic listed below.

Table 2 displays the levels of Prior Knowledge (PK) the students possessed on the text topics of focus. Notice the disparity in knowledge of the pre-Civil War topic (principles and facts relevant to the pre-Civil War period) and the knowledge of the immigration topics, topics less frequently treated in history courses. The absolute level of knowledge is also worthy of discussion. Even the pre-Civil War topic showed an average of only approximately 35% correct with a large variation in level of student knowledge. This average is comparable to other samples of students completing this task. For the immigration topics,

the percentages were dramatically lower, with Chinese Immigration knowledge at approximately 23% and General Immigration at about 19%. Since even a marginal answer is awarded one point, and there are twenty items, these results suggest that many of the students were leaving significant proportions of the items blank.

Table 2

Means and Standard Deviations of Preliminary Elements by Topic

Topic	Mean	*SD*	Total Possible
Prior Knowledge Civil War (CW)	28.32	10.62	80
Prior Knowledge Chinese Immigration (CI)	15.22	7.96	80
Prior Knowledge General Immigration (GI)	18.82	8.32	80
Literal Comprehension (CW)	7.86	2.76	14
Literal Comprehension (CI)	7.31	2.68	12
Literal Comprehension (GI)	6.40	2.73	12

The Prior Knowledge findings are consistent with earlier results (Baker & Clayton, 1989; Baker, Clayton, Aschbacher, Chang, & Ni, 1990; Baker, Niemi, Aschbacher, Ni, & Yamaguchi, 1991), showing that students had low levels of understanding of principles and events as measured by the prior knowledge tests. The reading comprehension scores were relatively low as well.

Means and standard deviations of student essay performance for each topic are presented in Table 3.

For these tables, a student's score was derived from averaging scores of all four raters. The absolute level of essay performance was low, particularly when subscore means, such as use of Prior Knowledge, are considered. For example, based on the General Content Quality (GCQ), students were writing inadequate explanations of the content. Students also showed high scores on the Text (TX) subscale, indicating that they depended heavily on the presented text for the ideas in their essay, a characteristic of less well prepared students. This finding is also supported by the low scores obtained on the use of Prior Knowledge in the essay, which suggests that students depended upon the text because they had no other reliable source of knowledge. Note that the order of Prior Knowledge proficiency—Civil War, Chinese Immigration, General

Immigration—is the same as reported in Table 2, on the prior knowledge test. The pattern of these results is comparable across the three topics, a finding that will be discussed later, and portrays relatively low levels of understanding.

Table 3

Means and Standard Deviations of Student Essay Performance for Each Topic by Subscale

	GCQ	PK	PN	TX	MI	A
			Civil War			
Mean[a]	2.48	1.69	1.83	3.11	2.11	2.50
SD	.72	.78	.57	.97	.89	.70
			Chinese Immigration			
Mean[a]	2.36	1.24	1.72	3.70	1.57	2.22
SD	.72	.36	.48	.84	.57	.66
			General Immigration			
Mean[a]	2.31	1.14	1.77	3.23	1.57	2.24
SD	.68	.48	.47	1.10	.57	.62

Note. Means and standard deviations are averaged across four raters. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

[a] 5 possible points.

Order and Classroom Effects

Analysis of variance was conducted to determine whether significant topic order (Chinese or General Immigration topics) or classroom effects existed. Recall that the immigration tasks were presented in counterbalanced order. No main effects at the .01 level were found for topic order or classroom membership.

Studies of Reliability

Classical reliability studies. Essay scores were analyzed to determine the level of interrater reliability using classical reliability techniques. Table 4 presents interrater reliability for all combinations of the four raters for each of the scoring dimensions of the essay rubric.

Table 4

Interrater Reliabilities (Alpha)

Raters				GCQ	PK	PN	TX	MI	A
						Civil War			
1	2	3	4	.87	.91	.72	.86	.70	.87
1	2	3		.84	.89	.65	.87	.57	.82
1	2		4	.84	.91	.78	.80	.70	.85
1		3	4	.84	.87	.53	.78	.70	.83
	2	3	4	.81	.87	.66	.81	.70	.82
1	2			.75	.87	.76	.85	.48	.77
1		3		.80	.81	.31	.76	.35	.71
1			4	.82	.87	.57	.62	.42	.85
	2	3		.77	.82	.50	.84	.53	.78
	2		4	.76	.88	.76	.64	.77	.73
		3	4	.71	.75	.40	.68	.51	.74
						Chinese Immigration			
1	2	3	4	.91	.79	.85	.89	.73	.87
1	2	3		.88	.69	.81	.85	.63	.81
1	2		4	.91	.86	.86	.88	.72	.85
1		3	4	.87	.67	.76	.83	.65	.85
	2	3	4	.88	.69	.78	.85	.67	.84
1	2			.87	.81	.87	.86	.60	.77
1		3		.79	.37	.63	.76	.45	.74
1			4	.87	.82	.74	.79	.65	.82
	2	3		.82	.51	.63	.75	.51	.71
	2		4	.87	.78	.80	.85	.66	.79
		3	4	.80	.43	.63	.76	.55	.82
						General Immigration			
1	2	3	4	.88	.80	.70	.90	.52	.84
1	2	3		.83	.70	.58	.89	.51	.80
1	2		4	.86	.81	.59	.84	.44	.82
1		3	4	.83	.75	.50	.85	.36	.80
	2	3	4	.86	.75	.77	.89	.46	.79
1	2			.82	.66	.51	.81	.54	.75
1		3		.71	.59	.23	.85	.35	.72
1			4	.77	.77	.20	.68	.00	.76
	2	3		.77	.58	.67	.86	.40	.73
	2		4	.82	.81	.70	.84	.34	.73
		3	4	.83	.62	.69	.82	.34	.71

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

These reliabilities suggest that raters are in agreement across topics on the elements of the scoring rubric, particularly the General Content Quality, Argumentation (A) (a newly added scale), and Text. Topic interacts with reliability on Prior Knowledge, and reliability coefficients are lowest for Misconceptions (MI). Interrater reliability was also analyzed in terms of exact agreement among pairs of raters on the score points and agreement within one score point. Tables 5-7 display the findings, showing that agreement is high between rater pairs within one score point. This result is undoubtedly affected by the skewed distribution of scores, resulting from relatively poor student performance.

Table 5

Percentage of Exact Agreement—Civil War (CW)

Raters				GCQ	PK	PN	TX	MI	A
				Percentage Exact Agreement					
1	2			54	68	59	51	27	44
1		3		59	62	35	32	40	46
1			4	59	62	51	37	40	62
	2	3		56	55	31	34	33	58
	2		4	67	75	58	30	40	63
		3	4	56	58	30	36	39	61
				Percentage Exact Agreement ±1 Score Point					
1	2			97	98	97	94	73	89
1		3		94	95	86	86	78	89
1			4	98	100	94	70	79	97
	2	3		97	95	94	91	67	97
	2		4	97	97	98	70	83	94
		3	4	92	95	91	75	67	94

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Table 6

Percentage of Exact Agreement—Chinese Immigration (CI)

Raters				GCQ	PK	PN	TX	MI	A
				Percentage Exact Agreement					
1	2			73	83	82	61	42	65
1		3		59	70	44	38	36	55
1			4	71	86	77	42	38	54
	2	3		68	73	38	42	53	67
	2		4	68	85	82	59	53	53
		3	4	55	76	44	55	64	55
				Percentage Exact Agreement ±1 Score Point					
1	2			100	100	100	96	91	100
1		3		99	99	99	85	89	94
1			4	97	100	97	92	92	95
	2	3		99	100	99	85	91	94
	2		4	100	100	99	92	91	99
		3	4	99	100	97	91	94	100

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Table 7

Percentage of Exact Agreement—General Immigration (GI)

Raters				GCQ	PK	PN	TX	MI	A
				Percentage Exact Agreement					
1	2			65	52	52	40	50	53
1		3		47	57	37	43	35	47
1			4	55	63	52	34	36	60
	2	3		67	68	60	50	40	67
	2		4	65	84	63	55	47	55
		3	4	53	65	48	45	68	57
				Percentage Exact Agreement ±1 Score Point					
1	2			97	97	97	71	86	95
1		3		90	90	88	77	80	93
1			4	95	98	92	65	81	95
	2	3		98	100	97	93	80	100
	2		4	98	100	100	95	89	97
		3	4	97	100	98	92	88	93

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Intrarater reliabilities based on 17 rescored essays were computed by subscale for each of the three topics and are reported below in Table 8. One rater showed considerable inconsistency (Rater 3) and the dimension of Prior Knowledge on the immigration topics showed unstable judgments. Yet, these findings suggest that, for the most part, individual raters are reasonably stable in their assessments of student papers.

Table 8

Intrarater Reliabilities

Raters	GCQ	PK	PN	TX	MI	A	n
			Civil War				
1	.89	.93	.72	.55	.17	.88	15
2	.82	.95	.92	.90	.75	.94	17
3	.87	.83	—[a]	.81	—[a]	.60	17
4	.92	.91	.85	.95	.96	.84	17
			Chinese Immigration				
1	.98	1.00	1.00	.88	.73	.96	15
2	.93	.74	1.00	.99	.98	1.00	15
3	.97	.87	1.00	1.00	.98	.97	15
4	.97	1.00	1.00	.99	1.00	.97	16
			General Immigration				
1	.89	.97	1.00	.99	.82	.94	15
2	1.00	.80	.80	.97	.86	1.00	15
3	.88	.69	.72	.86	.95	.73	16
4	.93	1.00	1.00	.95	1.00	.91	15

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

[a] Insufficient data.

Factor analytic rater reliability studies. Although factor analyses have been conducted to assist our revision of the scoring rubric (Baker, Clayton, Aschbacher, Chang, & Ni, 1990), with this dataset we conducted factor analyses studies to shed light on rater agreement (Tables 9-11) in order to influence the strategies we might use in rater selection and training. In these analyses, we see clear consistencies.

Table 9

Principal Component Factor Analysis, Percent of Variance Explained by the Four Raters on Each Dimension (Civil War)

	% Variance Explained
Dimension	Factor 1
GCQ	73.0
PK	79.6
PN	57.0
TX	70.8
MI	53.8
A	72.0

Table 10

Principal Component Factor Analysis, Percent of Variance Explained by the Four Raters on Each Dimension (Chinese Immigration)

	% Variance Explained
Dimension	Factor 1
GCQ	79.3
PK	63.1
PN	68.5
TX	75.1
MI	56.0
A	74.4

Table 11

Principal Component Factor Analysis, Percent of Variance Explained by the Four Raters on Each Dimension (General Immigration)

	% Variance Explained		
Dimension	Factor 1	Factor 2	Total
GCQ	74.8		
PK	65.3		
PN	53.9		
TX	78.8		
MI	41.4	25.6	67.0
A	69.1		

First, the General Content Quality scoring dimension is either first or second in the percent of consistency among raters (see Table 12). The Argumentation subscale also generated consistency among raters. Number of Principles and Misconceptions were subscales with the lowest consistency, and in fact, two factors were extracted for Misconceptions on the General Immigration topic. Although it is tempting to posit the limited variability of the scores to explain these findings, the data do not fully support this analysis. A better explanation seems to hinge upon the raters' and students' familiarity with the topics. For instance, the Civil War topic focused on the Lincoln-Douglas debates, a topic well known to history teachers and typically covered at some level in history courses. Thus raters are able to identify students' use of prior knowledge with great consistency. They understand what relevant prior knowledge is and because of their own topic familiarity can see it in student work. Similarly, these raters have a good idea of logical argument and integration of elements in the essay, accounting for their consistency on that dimension. On the other hand, on the immigration topics, topics not necessarily a common part of existing history curricula, raters showed less consistency in the Prior Knowledge subscale and much greater agreement (either the highest or second highest level) on the texts related to Chinese and General Immigration. This focus on the text suggests that raters in their rating, as students in their writing, depend upon the provided materials when their own prior knowledge base is weak.

Table 12

Order of Principal Components Explained by Raters Across Topics

	CW	CI	GI
GCQ	2	1	2
PK	1	5	4
PN	5	4	5
TX	4	2	1
MI	6	6	6
A	3	3	3

Note. CW = Civil War, CI = Chinese Immigration, GI = General Immigration, GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Rater and Topic Impact on Assessment Design

Generalizability Study

A generalizability study was conducted to determine rater reliability with the specific additional goal of determining the degree to which topics (Civil War, Chinese Immigration, and General Immigration) in our performance assessment were exchangeable. For each of the six scoring dimensions (General Content Quality, Prior Knowledge, Principles, Text, Misconceptions, and Argumentation) a two facet, completely crossed analysis of variance (ANOVA) design was used (see Tables A-F in Appendix E). In each of the ANOVA designs, topics (with three levels—Civil War, Chinese Immigration, and General Immigration) and raters (with four levels) were used as repeated measures facets.

The estimates of variance components were obtained assuming the topic and rater facets as random facets. Similar trends were obtained using random or fixed models for either facet. Table 13 presents estimates of variance components for the six dimensions of the scoring rubric. G-coefficients for scoring dimensions suggest that the General Content Quality and Argumentation dimensions show less variability over raters and topics than do other dimensions. The low estimated generalizability for the Misconceptions dimension, for example, is supported by the greatly variant intrarater reliability study. The variance components for subjects-by-topics and for subjects-by-topics-by-raters were relatively large for all dimensions. Variance components for all six dimensions for subjects-by-raters are considerably smaller than the subjects-by-topics effects. In general, the implications of the findings are that there is good rater consistency, and the contribution of adding additional raters is small. The variation in the topic facet is relatively high and particular scoring dimensions show greater sensitivity to topic. The Prior Knowledge and Misconceptions dimensions are particularly subject to topic variation. These findings are represented in a slightly different way by reviewing the Pearson correlation coefficients among topics. The correlation between the General Content Quality dimensions of the Civil War and General Immigration topics is .58; for the Argumentation dimension of the same topics, .49. (All correlations were computed using 68

subjects, compensating for missing data.) The correlations between the Civil War and Chinese Immigration topics on General Content Quality were .53, and .55 for the Argumentation scale. The relationship between the immigration tasks was, oddly, somewhat lower, with General Content Quality correlated .36 and Argumentation, .40.

Table 13

Estimates of Variance Components for the 6 Dimensions for 4 Raters and 3 Topics

	Dimensions					
Source of Variations	GCQ	PK	PN	TX	MI	A
Subjects	.225	.091	.068	.279	.027	.187
Topics	.005	.000	.038	.058	.073	.017
Raters	.001	.046	.006	.082	.024	.015
Subjects by Topics	.186	.091	.205	.500	.269	.161
Subjects by Raters	.037	.027	.034	.029	.055	.036
Topics by Raters	.012	.007	.015	.064	.054	.032
Subjects by Topics by Raters	.185	.219	.146	.419	.561	.193
G-Coefficient	.722	.433	.621	.572	.151	.703

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Table 14 presents data showing the impact of adding more topics and/or raters to the design of performance assessments. The g-coefficients were generated using results from all three topics. Table 14 shows the relative benefit for various combinations. Using two raters rather than one yields a worthwhile increase in reliability. Using three topics (the same amount of student time as in this study) and two raters, we obtain a g-coefficient of .67. These same data are portrayed in Figure 1. The figure shows the diminishing returns for additional raters, whereas additional topics have substantial effects on estimated generalizability. Except for the Misconceptions dimension, remaining scales will show the same general shape as in Figure 1. Results for other dimensions are provided in Tables 15-19. Findings for the Argumentation dimension (in Table 19) are strikingly similar, with a g-coefficient of .64 obtained from three topics and two raters. For the other dimensions, three topics and two raters yields lower estimated generalizability. In fact, except for Argumentation, for those dimensions it is

not possible to obtain acceptable levels of generalizability with reasonable numbers of topics and raters.

Table 14

Generalizability Coefficients for Various Combinations of Number of Topics and Number of Raters (GCQ) Based on CW, CI and GI Topics

	Rater				
Topic	1	2	3	4	5
1	.36	.43	.46	.48	.49
2	.50	.59	.62	.64	.65
3	.58	.67	.70	.72	.73
4	.63	.72	.75	.77	.78
5	.67	.75	.78	.80	.81

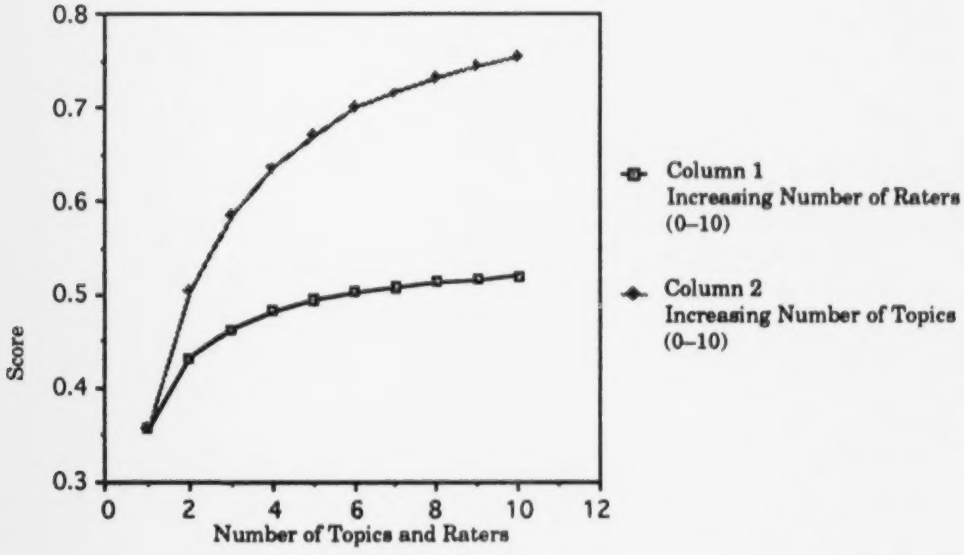

Figure 1. Score generalizability of General Content Quality scores of extended history tasks as a function of number of history topics and number of raters.

Table 15

Topics by Raters, Prior Knowledge (PK)

	Rater				
Topic	1	2	3	4	5
1	.15	.19	.20	.21	.22
2	.24	.30	.33	.34	.35
3	.31	.38	.41	.43	.44
4	.36	.44	.48	.50	.51
5	.39	.48	.52	.54	.56

Table 16

Topics by Raters, Principles/Concepts – Number (PN)

	Rater				
Topic	1	2	3	4	5
1	.21	.30	.34	.37	.39
2	.33	.44	.50	.53	.55
3	.41	.53	.59	.62	.64
4	.47	.59	.64	.68	.70
5	.51	.63	.68	.72	.74

Table 17

Topics by Raters, Proportion of Text Detail (TX)

	Rater				
Topic	1	2	3	4	5
1	.23	.28	.30	.31	.32
2	.36	.43	.46	.47	.48
3	.45	.53	.56	.57	.58
4	.52	.59	.62	.64	.65
5	.57	.64	.67	.69	.70

Table 18

Topics by Raters, Misconceptions (MI)

	Rater				
Topic	1	2	3	4	5
1	.03	.04	.05	.06	.06
2	.05	.08	.10	.11	.12
3	.07	.11	.14	.15	.16
4	.09	.14	.17	.19	.20
5	.11	.16	.20	.22	.23

Table 19

Topics by Raters, Argumentation (A)

	Rater				
Topic	1	2	3	4	5
1	.32	.40	.44	.46	.47
2	.47	.56	.60	.62	.64
3	.55	.64	.68	.70	.72
4	.60	.69	.73	.75	.77
5	.64	.73	.77	.79	.80

The dimensions of the scoring rubric are intended for different purposes. The General Content Quality (GCQ) dimension is the only dimension to be used for either public accountability or reporting purposes, while the remaining dimensions will be used for feedback to teachers and students. Thus, our focus is on refining the generalizability of the GCQ scale.

Because of the lower relationships found among the Chinese Immigration and General Immigration topics, we re-estimated g-coefficients using only data from the Civil War and the General Immigration tasks. Table 20 presents the findings that demonstrate a g-coefficient of .73 on the General Content Quality dimension for a three-topic, two-rater design. Thus, refining the topic design alone would improve the quality of measurement.

Table 20

Generalizability Coefficients for Various Combinations of Number of Topics and Number of Raters (GCQ) Based on CW and GI Topics

	Rater				
Topic	1	2	3	4	5
1	.41	.51	.56	.58	.60
2	.56	.66	.70	.73	.74
3	.64	.73	.77	.79	.80
4	.69	.78	.81	.83	.84
5	.72	.80	.84	.85	.86

Topic-by-Student Characteristics

One general issue of concern was whether certain topics, because of their content focus, would have a particularly motivating effect on students. In this study, for instance, two topics focused on immigration issues, one on immigration in general and one on Chinese immigration. We were particularly interested in determining whether ethnic status interacted with students' essay performance. Sixty-one students provided ethnic self-descriptions on our background questions. Forty-nine students were white, nine Asian, and nine African-American or Latino. The unanticipated disparity among ethnic distribution limits the generalizability of our findings. Male and female status was also coded and the sample consisted of 32 males and 29 females. Separate multivariate analyses of variance (MANOVA) were conducted for each of the three assessment topics, using a factorial design crossing gender and ethnic status for each of the six scoring dimensions. No significant effects for ethnicity were found. Figures 2-4 display significant gender effects found on the immigration topics for the General Content Quality, Principles and Argumentation dimensions ($p<.01$); on the Civil War topic, significant differences were found for Principles and Argumentation ($p<.05$) dimensions, and the General Content Quality dimension skirted significance ($p=.052$). These findings are consistent with other studies showing superior performance for females on writing tasks.

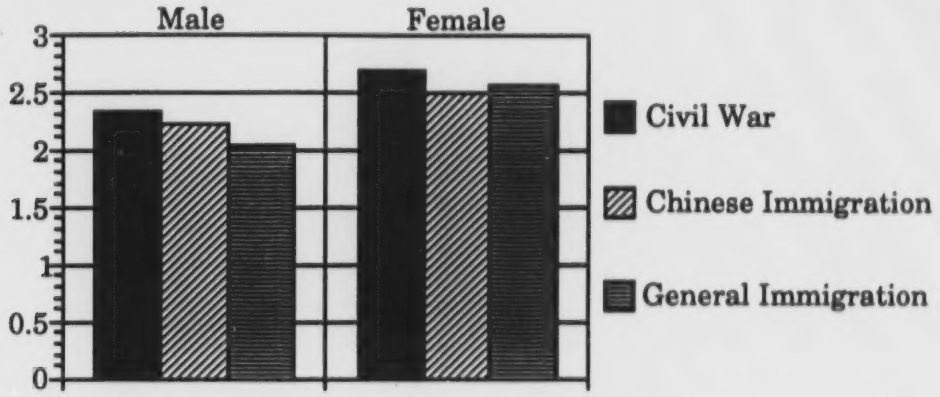

Figure 2. Gender effects on General Content Quality by topics.

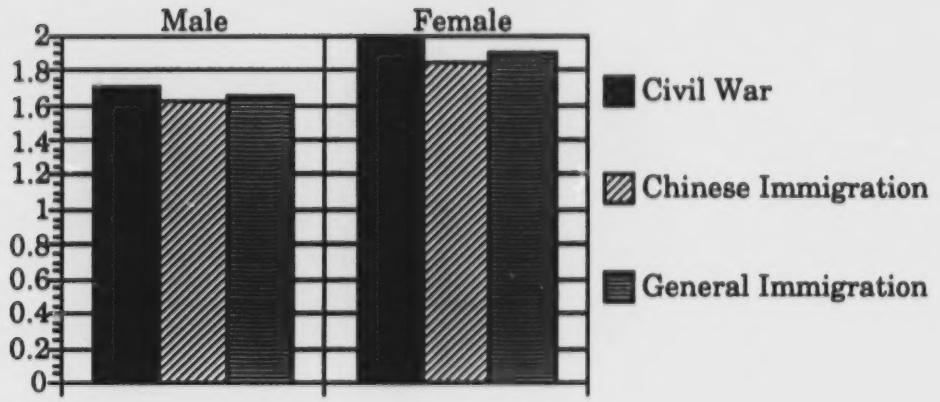

Figure 3. Gender effects on Principles by topics.

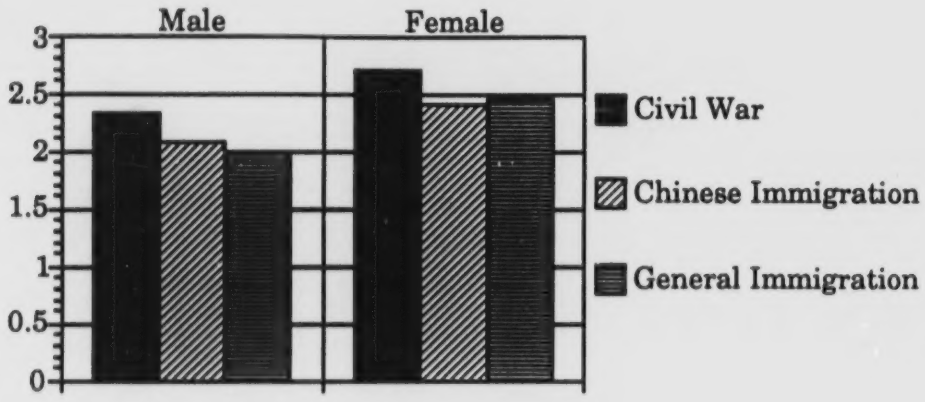

Figure 4. Gender effects on Argumentation by topics.

Relationship of Performance Measures to Measures of Test Anxiety, Effort, and Metacognition

Correlation coefficients were computed assessing the extent to which self-report measures of anxiety, student effort, and metacognitive processes were related to performance tasks. Two types of test anxiety measures were administered: (a) trait anxiety, a measure of the typical anxiety felt by the student in assessment settings; and (b) state worry, the particular concerns felt during the preceding task. State measures were administered immediately following the completion of each section of the performance task (prior knowledge, reading comprehension, and essay). The trait anxiety measure is comprised of Worry and Emotionality scales (Spielberger, 1983). These measures were found to correlate only with the state measures of Worry administered following each of the components of the assessment. State measures of Anxiety did not correlate with performance measures, an unusual finding since state anxiety usually tracks difficulty of achievement measures. An explanation for this finding comes from the observers, who reported a relatively low level of engagement and energy on the part of these students, compared with students who have completed the same tasks in other studies.

In contrast to these findings, a newly developed self-report Effort Scale (O'Neil & Snow, in preparation) was found to relate significantly and positively to the essay dimensions in this assessment study. For all three topics, Effort scores significantly correlated with General Content Quality and Argumentation dimensions in the range of .33 to .49, a relationship beyond the .01 probability level. For the Civil War and General Immigration topics, Effort correlated with Principles at the .05 level and for the Chinese Immigration topic at the .01 level. The two immigration topics also showed correlations with Effort for the Prior Knowledge dimension, beyond the .01 level for both topics. Responses to the Effort scale were intercorrelated across occasions, between .38 and .66, both beyond the .01 level of probability.

Relationship with Measures of Effort, Anxiety, and Metacognition

The metacognitive or "thinking" questionnaire consists of self-reports on four different five-item dimensions: Planning, Cognitive Strategies, Awareness, and Self-Checking. The data are reported in Table 21. Of the 24 possible correlations between the six scoring dimensions and the four subscales of the metacognitive measure, only one significant correlation was found (.28, $p<.05$) for Prior Knowledge and Self-Checking, very likely a chance event. For Chinese Immigration, significant relationships were found for all four subscales for General Content Quality. For the General Immigration tasks, relationships were found for Prior Knowledge and Cognitive Strategies, Awareness, and Self-Checking, and for Argumentation between Planning and Cognitive Strategies. Self-Checking correlated significantly (at the .05 level) with Misconceptions, a finding that would require some explanation if replicated in other studies. Given the low levels of knowledge and performance on both of these topics, we might explain the expected relationships in terms of task difficulty. Metacognitive skills are really only invoked for truly difficult and/or unfamiliar tasks.

Table 21

Significant Correlations Between Metacognitive Scales and Essay Performance Dimensions

Dimension	Planning	Cognitive Strategies	Awareness	Self-Checking
		Civil War		
GCQ				
PK				.28*
PN				
TX				
MI				
A				
		Chinese Immigration		
GCQ	.31*	.42**	.45**	.27*
PK				
PN				
TX				
MI				
A	.29*	.39**		
		General Immigration		
GCQ				
PK		.28*	.37*	.31*
PN				
TX				
MI				.29*
A	.29*	.28*		

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

*$p<.05$. **$p<.01$.

Dimensionality of Scoring Rubric

The development of the essay scoring scheme derived from the analysis of essays developed by experts, that is, individuals with advanced degrees in American history. The experts organized their explanations using major principles or concepts which they fleshed out with specific knowledge about the historical period or the authors of the texts. In the same study, novices—high school students—wrote essays that depended to a great degree on the texts provided and exhibited major misconceptions as well as mistakes of fact.

Factor analysis was conducted on this dataset to study the dimensionality of the scoring rubric. The findings, reported in Tables 22-24 suggest a two factor solution, as predicted by the original design of the scoring rubric. One factor includes General Content Quality, Principles, and Argumentation. Prior Knowledge is appropriately loaded on the Civil War topic, but on the less familiar topic of General Immigration, Prior Knowledge loads on both factors. This is a factor measuring expertise. The second factor involves Text dependence and Misconceptions. These findings support the theory that deep understanding consists of the integration of high level and specific prior knowledge with newly acquired information.

Table 22

Rotated Factor Matrix for Essay Dimensions—Civil War

Dimension	Factor 1	Factor 2
GCQ	.97316	-.03063
PK	.66416	-.51764
PN	.87470	-.28833
TX	.14128	.94054
MI	-.23751	.42313
A	.95558	.02748

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Table 23

Rotated Factor Matrix for Essay Dimensions—Chinese Immigration

Dimension	Factor 1	Factor 2
GCQ	.88853	.35478
PK	.76095	-.39672
PN	.91099	.15188
TX	.05208	.88183
MI	-.15797	-.59111
A	.90246	.32417

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

Table 24

Rotated Factor Matrix for Essay Dimensions—General Immigration

Dimension	Factor 1	Factor 2
GCQ	.96575	-.00390
PK	.54751	.72349
PN	.91625	-.16531
TX	.50323	-.68239
MI	-.05963	.59391
A	.97368	-.00465

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation.

The level of correlations among the variables—Prior Knowledge, Principles, and Argumentation—with the General Content Quality is remarkably consistent across topic areas. However, the relationships among these variables and the dimensions of Text and Misconceptions change in the light of the raters' and students' differential understanding of particular topics. For example, dependency upon the text is uncorrelated with General Content Quality in the Civil War topic, based upon one Lincoln-Douglas debate. This is what the model would predict, for experts may include some text relevant information in their answers. But on the General Immigration topic, text use is positively correlated with General Content Quality. A reasonable interpretation of these findings relates to the lack of experience of both students and raters with immigration content. When raters have little knowledge of the subject matter, they will tend to value information derived from the texts used in the assessment because they have no other basis to judge students efforts.

This finding also points up an additional issue in the use of performance assessment. Performance assessment requires cognitive demands that are not often made of students in school. Previous studies (Baker, Freeman, & Clayton, 1991; Baker, Niemi, Gearhart, & Herman, 1990) showed that students reported little if any experience in multistep performance tasks such as those provided. Current reforms in curriculum frameworks will be expected to add new content, such as the immigration topics included in this study. In these

cases, neither student, teacher, nor rater may be very well informed or competent in the subject matter. One might expect different relationships among scoring dimensions between instructed and uninstructed students as rated by well informed or less informed teachers. We must be cautious in making inferences from tasks where instructional history is known to be scant.

Correlations of the initial prior knowledge tests, reading comprehension scores, and essay dimensions were computed and the correlation matrix is presented in Tables 25-27 below.

Table 25

Correlations Among Initial Prior Knowledge Test and Essay Dimensions by Topic—Civil War

	GCQ	PK	PN	TX	MI	A	RC
Prior Knowledge Test	.49**	.30*	.41**	.07	-.24	.50	.50**
Reading Comprehension	.41**	.23	.38**	.00	-.31*	.45**	

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation, RC = Reading Comprehension.

*$p<.05$. **$p<.01$

Table 26

Correlations Among Initial Prior Knowledge Test and Essay Dimensions by Topic—Chinese Immigration

	GCQ	PK	PN	TX	MI	A	RC
Prior Knowledge Test	.42**	.36**	.40**	.01	-.11	.44	.54**
Reading Comprehension	.51**	.32**	.57**	.18	-.19	.53**	

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation, RC = Reading Comprehension.

*$p<.05$. **$p<.01$

Table 27

Correlations Among Initial Prior Knowledge Test and Essay Dimensions by Topic—General Immigration

	GCQ	PK	PN	TX	MI	A	RC
Prior Knowledge Test	.28*	.33*	.21	.01	.12	.26*	.40**
Reading Comprehension	.47**	.13	.44**	.31	-.02	.45**	

Note. GCQ = General Content Quality, PK = Prior Knowledge, PN = Principles/ Concepts-Number, TX = Text, MI = Miscellaneous, A = Argumentation, RC = Reading Comprehension.

*$p<.05$. **$p<.01$

These relationships are remarkably consistent. Weaker relationships are likely a function of lack of knowledge and lack of variability.

Relationship of Performance Tasks and Grade Point Average

To assess external validity of these performance tasks, we conducted a correlational analysis of students' grade point average (GPA) and performance measures in the study. Scores for each student's performance were averaged over all three topics. Two different GPA measures were used. The first was based on all possible English and history classes the students might have taken, with the exception of American history courses. The second GPA was computed using grades from the two available American history courses.

With the first type of GPA, General Content Quality and Argumentation dimensions correlated at .45 and .46 respectively with GPA. These correlations were significant beyond the .01 level of probability. These two dimensions are highly intercorrelated (.92) and probably measure students' ability to access and organize relevant knowledge for the particular task. The Principles dimension was correlated with GPA at .39, significant beyond the .05 level. The prior knowledge test scores correlated with GPA for the three topics at 55, .36, and .52 respectively, all beyond the .01 level of probability. These correlations are surprising in light of the lack of variability on the performance measures.

One would expect, however, the validity of these history performance tasks to be judged in terms of performance in American history courses. We

repeated the correlational analysis using GPA based on American history course grades. This analysis generated results that strongly support the relevance of the scoring dimensions to other indices of history performance. Significant relationships were found for General Content Quality (r=.44), Principles (r=.54), and Argumentation, (r=.47), all beyond the .01 probability level. Misconceptions correlated negatively and significantly (r=-.28), beyond the .05 probability level. The lack of correlation with the Prior Knowledge essay dimension may be a result of the rarity with which immigration topics are taught in American history courses and students' inability to apply such knowledge. Prior knowledge test performance correlated less strongly with American history courses, from .41 to .51, but still beyond the .01 level of significance. The history course GPA also correlated negatively and significantly (beyond the .05 level) with the trait anxiety scales of worry (r=.30) and emotionality (r=.30) as well as state anxiety measures (r=.29).

Interpretation of Findings

This research has generated a number of significant findings. They may be classified in four major categories: (a) task structure; (b) topic variability; (c) rater reliability; (d) external validity.

Task Structure

Theory confirmation. The findings replicate earlier results indicating that the scoring rubric is appropriate and generalizable across topics. The factor analytic studies replicate the validation of the expert scoring dimensions with a two factor solution differentiating between expert and novice performance characteristics. Specifically, the finding is that deep understanding consists of activating relevant prior knowledge, structuring the knowledge into major concepts or principles, selecting relevant details for elaboration from prior or provided knowledge, and logically relating information. This finding is extremely important because it suggests that common scoring dimensions, based on constructivist theories of learning, can be practically applied to different topics. Instructionally, the implications of these findings are that teachers can focus their attention on common, presumably learnable dimensions of subject matter understanding and

develop instructional programs that have coherence within subject matter across grades.

The cost considerations are no less important for managers of assessment—for our data suggest that the cost of developing and training on separate scoring schemes for different topics is not warranted. This finding will permit greater interpretability of data across testing occasions as well.

Practical implications. The fact that scoring dimensions such as Argumentation, Prior Knowledge, and Principles correlate highly with the estimate of General Content Quality fits the formulation of constructivist theory. It also presents a set of decision options related to both the particular purpose of the measurement and resource constraints. Where components of a rubric are intercorrelated, we might argue that it is unnecessary to score all components. If we were in a purely group or program accountability situation—for example, some state assessment programs—we might report simply the General Content Quality score to parents, policymakers, and the public as a measure of the deep understanding of subject matter. If these performance tasks were to be used exclusively to improve classroom instruction, the detailed features of the scoring rubric should be retained, scored by teachers, to provide insights to both students and teachers about needed instructional directions. The view that performance assessment contributes to teachers' subject matter understanding would support the retention of all rubric dimensions. But, as we are coming to understand, only rarely are performance tasks used exclusively for one or another purpose. Except for selection examinations, most test performances are thought to be the product of student ability, effort, and school experience. Thus, retaining scoring along all rubric dimensions, both the positive (e.g., Principles) and the negative (e.g., Misconceptions), would provide the most guidance for instruction and cues to permit the improvement of performance. However, more scoring dimensions equal more cost.

If costs were a major consideration, retaining task components with the best (to date) technical characteristics would be recommended and would result in the use of the 20-item prior knowledge test and the General Content Quality score for the essay. At this point, it is difficult to say whether the Principles or Prior Knowledge essay dimensions should similarly be retained,

particularly because of the students' lack of instructional experience with the topics studied. The reliabilities of these dimensions varied with the particular topics. Studies are now underway to assess performance under experimental instructional conditions. Comparing directly instructed and uninstructed students on specifically taught and transfer topics will shed light on the utility of the fully elaborated scoring dimension.

Our data suggest that the retention of the reading comprehension task component is unwarranted, and we intend to drop this component from our task.

Measures of anxiety, effort, and metacognition. Our tasks included measures of anxiety, effort, and metacognition. Should these be routinely included in assessments?

Test anxiety relationships in this particular task did not replicate the usual or previously found significant, negative relationships (Baker & Clayton, 1989). The students were reported by our observers to be relaxed during the data collection. Whether this was out of ennui or the comfort of a more "authentic" task we do not know. For future research, we will continue this component of data collection.

We intend to continue to include the effort scale in our studies of performance, partly as a strategy to cross validate both sets of measures, as well to provide an independent measure of effort for review by teachers and policymakers. The findings on the metacognitive scale, with predicted correlations found for only the most difficult task, must be interpreted with care, although it is plausible that metacognitive activity is precisely most functional for very difficult tasks. Until our findings are clearer, we plan to keep this measure in our data collection.

Topic Variability

The bane of performance assessment has been the idiosyncratic performance of individuals on different tasks so that a reliable estimate across a family of tasks requires excessive time and testing resources. The use of task specifications, resulting in a common task structure and common scoring rubric, controls some of the variation associated with performance tasks. The correlations between topics for specification-developed performance

assessments were .58 and .53 for correlations involving the previously field tested Civil War topic with each of the immigration topics and .37 for the immigration pair for the General Content Quality dimension of the essay score, and between .60 and .71 for the prior knowledge tests. These results compare favorably with the free-response section of the Advance Placement Examinations—with correlations between .28 and .35 between pairs of items—(Bleistein, Whitehurst, & Flesher, 1991). The College Board for Advanced Placement Examinations reports score reliability of .50 for two tasks and two raters. In studies of clinical tests administered to medical students, Julian and Schumacher (1988) report intertask correlations of .27 and .05 for scores on simulation-based performance tests. Results from studies of the bar exam (Klein, 1991) show correlations among essay tasks in the .22-.44 range and correlations of .37 among performance tasks.

The results of the g-study suggest that given the high level of rater reliability, a design with three tasks and two raters would result in a g-coefficient of approximately .72 for the reading-essay task component (approximately 70 minutes a student). This finding is surpassed only by Swartz, Patience, and Whitney (1985) on the GED test of writing skills, cited in a review by Dunbar, Koretz, and Hoover (1991). Given the complexity of the history scoring rubric (assessing various types of knowledge sources and applications) compared to typical writing assessment rubrics focusing on organization, style, and expression, our results are extremely encouraging.

This finding makes it more likely that performance tasks in the multi-step model provided can be feasibly used in assessment. One way to think about the problem is in terms of the hours of student time necessary to achieve various levels of reliability among tasks. In our studies the actual time for assessment was a sum of the prior knowledge test (10 minutes), the time to read the texts (20-30 minutes), and the essay section (50 minutes), for a total of approximately 1 1/2 hours a task. It is reported for the bar examination that six hours of testing would yield about a .69 reliability score for essays and a .48 reliability for performance tests. In our data, we exceed the .48 level in one hour and the .69 level in less than three hours of student essay writing (using a two-rater model). It may be an oversimplification, but using parallel task design structures appears to double our assessment efficiency for the same

technical quality. This efficiency, for assessments administered in school environments, preserves valuable time for instruction.

Our findings may be also considered in the light of the low levels of understanding of our topics by our students (and raters). Unlike persons taking the GED, clinical medical examinations, the bar examination, or the Advanced Placement history examinations, our students were unprepared and apparently somewhat unmotivated. We look forward to the results of an on-going study to assess the validity of these measures on instructed students.

Rater Reliability

Our rater reliability is generally satisfactory for, as the g-studies demonstrate, using more than two raters adds relatively little precision to the reliability of the assessment. High rater reliability for all three topics (Civil War, Chinese Immigration, and General Immigration) was achieved on the General Content Quality dimension for four raters (r=.87, .91, and .88 respectively). Reliability coefficients were also obtained for Prior Knowledge (.91, .79, and .80 respectively), Text (.86, .89, and .90) and Argumentation (.87, .87, and .84). Although Principles (.72, .85, and .70, respectively) and Misconceptions (.70, .73, and .52) were lower, effort to raise reliability of the Principles dimension is warranted because of its strong relationship with the General Content Quality dimension.

Rater reliability has been achieved across topics using the same scoring dimensions and the same training regimens (described in Baker, Niemi, Aschbacher, Ni, & Yamaguchi, 1991). Essentially, raters are trained not only to agree with one another but to rate papers so to be consistent with pre-scored papers exemplifying desired validity characteristics. Secondly, rater fidelity to the rubric is checked and recalibrated (through training) during the scoring sessions. Clearly, the investment in training raters and maintaining rubric fidelity is worthy of continued resources, particularly when analyzed from the cost perspective—that is, the cost of adding multiple raters. In large scale assessment, one can estimate a cost per paper per rating of approximately $2.60. Clearly, single ratings of papers will be sufficient for most large-scale assessment purposes. For assessments designed to focus on individual student advancement, double-rating for papers falling within boundaries of

the cut-score would be desirable, and rater training would need to emphasize distinctions between pass, fail, and honors boundaries.

Factor analytic studies suggest the importance of understanding the degree to which mediating variables influence raters' scores. Rater background knowledge is one area which we intend to pursue in our research. Secondly, we are also interested in learning more about the differential cognitive demands of various scoring rubrics. We are pleased with our findings because our reading of the literature in cognitive task analysis (Glaser, Lesgold, & Gott, 1991), suggests our raters are accomplishing sophisticated processes: the determination of sources of knowledge, the determination of degree of knowledge and elaboration, the determination of interrelationships among levels and sources of knowledge, and the rapid internalization of these elements into their own executed procedures. Thus, our achieving high levels of interrater reliability is a strong accomplishment—it is not simply agreement among raters on the best to worst in a particular distribution. It should be noted that, for the most part, our results with raters have replicated over four other independent data collections, where we used some of the same and some different topics, students with different levels of ability, and different age levels. Sample sizes have ranged from 60 to 320 subjects.

One concern with rating of performance tasks is that of bias. While we found significant differences favoring females in this small sample, this set of all white raters did not differentially rate students' performance based on their ethnic or racial type.

External Validity

In this study, only limited evidence of external validity was sought, and focused on grade point average. Findings confirmed our other results, that is, grade point average strongly correlates with our prior knowledge measures and certain of the essay dimensions in predicted directions. Correlations with an effort scale were also in predicted directions.

Summary

What Have We Learned From Our Research on History Performance Assessment?

1. We have developed a valid scoring scheme for assessing deep understanding. It supports the theoretical model under which it was developed in this replication. It is generalizable across history topics. It is ready for dissemination.

2. We have developed rater training procedures that produce reliable and valid scoring of student tasks in a limited period. The scoring rubric makes strong cognitive demands of the raters. The training procedure is ready for dissemination.

3. We have a task structure that reduces score variability such that fewer topics can be used to derive reliable scores for individual students. This technique is more efficient than found in most comparable studies. These relationships are all the more startling because of the lack of preparation and motivation among our students.

4. We have distinguished between assessment purposes and the utility of overall score and subscores.

5. We have found gender differences in this small sample, favoring females.

6. We have found supportive data for the validity of our measures in grade point average (GPA) and a scale measuring student effort.

7. We have systematically addressed validity criteria (Linn, Baker, & Dunbar, 1991) in our research studies: the criteria addressed include fairness, generalizability, cognitive complexity, content quality, reliability, cost and efficiency. We are in the process of conducting studies of transfer and designing research to assess the meaningfulness of tasks to students.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Aschbacher, P.R. (1991). *Alternative assessment: State activity, interest, and concerns* (CSE Tech. Rep. 322). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E.L. (1990). *New approaches to student assessment*. Invited presentation at the conference on Education Reform sponsored by the Rockefeller Foundation, Oakland, CA.

Baker, E.L. (1991). *Foundation for Teaching Economics evaluation report*. Los Angeles: Author.

Baker, E.L., Aschbacher, P., Niemi, D., Chang, S.C., Weinstock, M., & Herl, H. (1991). *Validating measures of deep understanding of history*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Baker, E.L., & Clayton, S. (1989). *The relationship of anxiety and performance in content-based higher order thinking*. Paper presented at the 10th Annual Conference of the Society for Test Anxiety Research, Amsterdam, The Netherlands.

Baker, E.L., Clayton, S., Aschbacher, P., Chang, S., & Ni, Y. (1990). *Measuring deep understanding of history: The integration of prior knowledge and knowledge acquisition in explanations*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition*. Englewood Cliffs, NJ: Prentice-Hall.

Baker, E.L., & Herman, J.L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement, 20*, 149-164.

Baker, E.L., & Linn, R.L. (1990). *Advancing educational quality through learning-based assessment, evaluation, and testing*. Institutional Grant Proposal for OERI Center on Assessment, Evaluation, and Testing. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Baker, E.L., Niemi, D., Aschbacher, P., Ni, Y., & Yamaguchi, E. (1991). Using cognitively sensitive assessments of history. In E.L. Baker (Ed.), *Designing and scoring content assessments in American history*. Los Angeles: University of California, Center for the Study of Evaluation.

Baker, E.L., Niemi, D., Gearhart, M., & Herman, J.L. (1990). *Validating a hypermedia measure of knowledge representation*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Baker, E.L., Niemi, D., Novak, J., & Herl, H. (in press). Hypertext as a strategy for teaching and assessing knowledge representation. In S. Dijkstra (Ed.), *Instructional models in computer-based learning environments*. Berlin: Springer-Verlag.

Baker, E.L., O'Neil, H.F., Jr., & Linn, R.L. (1991). *Policy and validity prospects for performance-based assessment*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Bleistein, C.A., Whitehurst, C.M., & Flesher, R.B. (1991). *College Board Advanced Placement Examination European History 3NBP* (Report No. SR-91-104). Princeton, NJ: Educational Testing Service.

California State Department of Education. (1988). *History–social science framework for California public schools*. Sacramento, CA: Bureau of Publications Sales.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education*, *4*(4), 289-303.

Glaser, R., Lesgold, A., & Gott, S. (1991). Implications of cognitive psychology for measuring job performance. In A.K. Wigdor & B.F. Green, Jr. (Eds.), *Performance assessment for the workplace. Volume II: Technical issues*. Washington, DC: National Academy Press.

Hively, W., Patterson, H., & Page, S. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275-290.

Julian, E.R., & Schumacher, C.F. (1988). *CBT pilot examination: Results and characteristics of CBX*. National Board of Medical Examiners Conference on Computer-Based Testing in Medical Education and Evaluation, Philadelphia, PA.

Klein, S.P. (1991). *The effect of content area and test type on bar exam scores* (unpublished draft report). Los Angeles: Author.

Linn, R.L. (1991). *Technical considerations in the proposed nationwide assessment system for the National Education Goals Panel.* Paper prepared for the Technical Planning Subgroup of the Goal 3 Resource Group to the National Education Goals Panel. Washington, DC: National Education Goals Panel.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1992, Winter). Complex, performance-based assessment: Expectations and validation criteria. *Evaluation Comment*, pp. 2-9.

Linn, R.L., Kiplinger, V.L., Chapman, C.W., & LeMahieu, P.G. (in press). Cross-state comparability of judgments of student writing: Results from the New Standards Project. *Applied Measurement in Education.*

Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Morris, L.W., Davis, M.A., & Hutchings, C.H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology, 73*, 541-555.

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington: U.S. Government Printing Office.

O'Neil, H.F., Jr., Baker, E.L., Jacoby, A., Ni, Y., & Wittrock, M. (1990). *Human benchmarking studies of expert systems.* Los Angeles: University of California, Center for Technology Assessment.

O'Neil, H.F., Jr., & Snow, R.E. (in preparation). *State effort scale.*

Popham, W.J. (1987). The merits of measurement driven instruction. *Phi Delta Kappan, 68*(9), 679-682.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction.* Boston: Kluwer Academic.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*, 2-16.

Sizer, T.R. (1992). *Horace's school: Redesigning the American high school.* Boston: Houghton-Mifflin.

Spielberger, C.D. (1980). *Manual for the Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C.D. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.

Swartz, R., Patience, W., & Whitney, D.R. (1985). *Adding an essay to the GED writing skills test: Reliability and validity issues* (GED Testing Service Research Studies, No. 7). Washington, DC: American Council on Education.

U.S. Department of Education. (1991). *America 2000: An education strategy*. Washington: Author.

U.S. Department of Labor. (1991). *What work requires of schools: A SCANS report for America 2000*. Washington, DC: Department of Labor, The Secretary's Commission on Achieving Necessary Skills.

Wolf, D.P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership*, *46*(7), 35-39.

Appendix A

Content Assessment Materials: Civil War

UCLA Content Assessment Project
(Field Testing, July, 1991)

Student Demographic Form

Student's Name _____

Student's ID _____

Birthday _____

Teacher's Name _____

School in 1990-1991 _____

Grade Level in Sept. 1991 _____

Please circle the appropriate number:

Gender:
- Male 1
- Female 2

Ethnicity:
- White 1
- Black 2
- Hispanic 3
- American Indian 4
- Asian/Pacific Islander 5
- Other 8

Name: _____

How Much Do You Know About U.S. History?

Directions: This is a list of terms related to U.S. History. Most of them are related to the period of the Civil War, but some of them are from other periods in U.S. History.

In the space after each term, write down what comes to mind when you think of that term in the context of U.S. History. A brief definition would be acceptable, or a brief explanation of why that person, place, or thing was important. If the term is general, such as "Civil rights," give both a general definition and a specific example of how the term fits into U.S. History, if you can.

Good Example: CIVIL RIGHTS. Rights guaranteed to all citizens regardless of *race, sex, religion, etc. Blacks fought for their civil rights in the 1960s. Martin Luther King, Montgomery bus boycott.*

Do not define the term by simply restating the same words.

Bad Example: SURVIVAL OF THE FITTEST. Only the fittest survive.

Even if you are not sure about your answer, but think you know something, feel free to guess.

There are probably more items here than you will be able to answer in the time given. Start with the ones you know best, and work quickly so that you can answer as many as possible. Then go back and answer the ones of which you are less sure. Do not spend too much time on one specific item.

1. popular sovereignty

_____

_____

2. Dred Scott

_____

_____

3. Communism

_____

_____

4. Missouri Compromise

5. industrialization

6. Gold Rush

7. bleeding Kansas

8. states' rights

9. Federalism

10. underground railroad

11. imperialism

12. Whig

13. Kansas-Nebraska Act

14. Abolitionists

15. sectionalism

16. westward movement

17. constitutionality

18. New Deal

19. party platform

20. balance of power

Name _____

Attitude Measure
(Prior knowledge)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on t his task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name:_____

LINCOLN - DOUGLAS DEBATE*

Directions: As Abraham Lincoln and Stephen Douglas campaigned for the office of Senator from the state of Illinois, they held seven joint debates throughout the state. Read the following passages to understand as well as possible what Lincoln and Douglas discussed in one of their debates.

STEPHEN A. DOUGLAS

Mr. Lincoln tells you, in his speech made at Springfield, before the Convention which gave him his unanimous nomination, that—

"A house divided against itself cannot stand."

"I believe this government cannot endure permanently, half slave and half free."

"I do not expect the Union to be dissolved, I don't expect the house to fall; but I do expect it will cease to be divided."

"It will become all one thing or all the other."

That is the fundamental principle upon which he sets out in this campaign. Well, I do not suppose you will believe one word of it when you come to examine it carefully, and see its consequences. Although the Republic has existed from 1789 to this day, divided into Free States and Slave States, yet we are told that in the future it cannot endure unless they shall become all free or all slave. For that reason, he says, that they must be all free. He wishes to go to the Senate of the United States in order to carry out that line of public policy, which will compel all the States in the South to become free. How is he going to do it? Has Congress any power over the subject of slavery in Kentucky, or Virginia, or any other State of this Union? You convince the South that they

* From *Political Debates Between Abraham Lincoln and Stephen A. Douglas* (Cleveland, 1902), pp. 43-47.

must either establish slavery in Illinois, and in every other Free State, or submit to its abolition in every Southern State, and you invite them to make a warfare upon the Northern States in order to establish slavery, for the sake of perpetuating it at home. Thus, Mr. Lincoln invites, by his proposition, a war of sections, a war between Illinois and Kentucky, a war between the Free States and the Slave States, a war between the North and the South, for the purpose of either exterminating slavery in every Southern State, or planting it in every Northern State. He tells you that the safety of this Republic, that the existence of this Union, depends upon that warfare being carried on until one section or the other shall be entirely subdued. The States must all be free or slave, for a house divided against itself cannot stand. That is Mr. Lincoln's argument upon that question. My friends, is it possible to preserve peace between the North and the South if such a doctrine shall prevail in either section of the Union? Each of these States is sovereign under the Constitution; and if we wish to preserve our liberties, the reserved rights and sovereignty of each and every State must be maintained. I have said on a former occasion, and I here repeat, that it is neither desirable nor possible to establish uniformity in the local and domestic institutions of all the States of this Confederacy. And why? Because the Constitution of the United States rests upon the right of every State to decide all its local and domestic institutions for itself. It is not possible, therefore, to make them conform to each other, unless we subvert the Constitution of the United States. Our safety, our liberty, depends upon preserving the Constitution of the United States as our fathers made it, inviolate, at the same time maintaining the reserved rights and the sovereignty of each State over its local and domestic institutions, against Federal authority, or any outside interference.

The difference between Mr. Lincoln and myself upon this point is, that he goes for a combination of the Northern States, or the organization of a sectional political party in the Free States, to make war until they shall all be subdued, and made to conform to such rules as the North shall dictate to them. His answer to this point, which I have been arguing, is, that he never did mean, and that I ought to know that he never intended to convey the idea, that he wished the people of the Free States to enter into the Southern States and interfere with slavery. Well, I never did suppose that he ever dreamed of entering into Kentucky to make war upon her institutions; nor will any Abolitionist ever enter into Kentucky to wage such war. Their mode of making war is not to enter into those States where slavery exists, and there interfere, and render themselves responsible for the consequences. Oh, no! They stand on this side of the Ohio River and shoot across. They stand in Bloomington, and shake their fists at the people of Lexington; they threaten South Carolina from Chicago. And they call that bravery! But they are very particular, as Mr. Lincoln says, not to enter into those States for the purpose of interfering with the institution of slavery there. I am not only opposed to entering into the Slave States, for the purpose of interfering with their institutions, but I am opposed to a sectional agitation to control the institutions of other States. I am opposed to organizing a sectional party, which appeals to Northern pride, and Northern passion and prejudice, against Southern institutions, thus stirring up ill-feeling and hot blood between brethren of the same Republic.

I ask Mr. Lincoln how it is that he proposes ultimately to bring about this uniformity in each and all the States of the Union. Does he intend to introduce a bill to abolish slavery in Kentucky? How is he to accomplish what he professes must be done in order to save the Union? There is but one possible

mode which I can see, and perhaps Mr. Lincoln intends to pursue it; that is, to introduce a proposition into the Senate to change the Constitution of the United States, in order that all the State Legislatures may be abolished, State sovereignty blotted out, and the power conferred upon Congress to make local laws and establish the domestic institutions and police regulations uniformly throughout the United States. Whenever you shall have blotted out the State Legislatures, and consolidated all the power in the Federal Government, you will have established a consolidated empire as destructive to the liberties of the people and the rights of the citizen as that of Austria, or Russia, or any other despotism that rests upon the necks of the people.

There is but one possible way in which slavery can be abolished, and that is by leaving a State, according to the principle of the Kansas-Nebraska bill, perfectly free to form and regulate its institutions in its own way. That was the principle upon which this Republic was founded, and it is under the operation of that principle that we have been able to preserve the Union thus far. Under its operations, slavery disappeared from New Hampshire, from Rhode Island, from Connecticut, from New York, from New Jersey, from Pennsylvania, from six of the twelve original slaveholding States; and this gradual system of emancipation went on quietly, peacefully, and steadily, so long as we in the free States minded our own business and left our neighbors alone. But the moment the abolition societies were organized throughout the North, preaching a violent crusade against slavery in the Southern States, this combination necessarily caused a counter-combination in the South, and a sectional line was drawn which was a barrier to any further emancipation. Bear in mind that emancipation has not taken place in any one State since the Free-soil party was organized as a political party in this country. And yet Mr.

Lincoln, in view of these historical facts, proposes to keep up his electional agitation, band all the Northern States together in one political party, elect a President by Northern votes alone, and then, of course, make a cabinet composed of Northern men, and administer the government by Northern men only, denying all the Southern States of this Union any participation in the administration of their affairs whatsoever.

ABRAHAM LINCOLN

Judge Douglas made two points upon my recent speech at Springfield. He says they are to be the issues of this campaign. The first one of these points he bases upon the language in a speech which I delivered at Springfield which I believe I can quote correctly from memory. I said there that "we are now far into the fifth year since a policy was instituted for the avowed object, and with the confident promise, of putting an end to slavery agitation; under the operation of that policy, that agitation had not only not ceased, but had constantly augmented." "I believe it will not cease until a crisis shall have been reached and passed. 'A house divided against itself cannot stand.' I believe this Government cannot endure permanently, half slave and half free." "I do not expect the Union to be dissolved" — I am quoting from my speech — "I do not expect the house to fall, but I do expect it will cease to be divided. It will become all one thing or the other. Either the opponents of slavery will arrest the spread of it and place it where the public mind shall rest, in the belief that it is in the course of ultimate extinction, or its advocates will push it forward until it shall become alike lawful in all the States, North as well as South."

What is the paragraph? In this paragraph, Judge Douglas thinks he discovers great political heresy. I want your attention particularly to what he has inferred from it. He says I am in favor of making all the States of this Union uniform in all their internal regulations; that in all their domestic concerns I am in favor of making them entirely uniform. He says that I am in favor of making war by the North upon the South for the extinction of slavery; that I am also in favor of inviting (as he expresses it) the South to a war upon

the North for the purpose of nationalizing slavery. Now, it is singular enough, if you will carefully read that passage over, that I did not say that I was in favor of anything in it. I only said what I expected would take place. I made prediction only — it may have been a foolish one, perhaps. I did not even say that I desired that slavery should be put in course of ultimate extinction. I do say so now, however, so there need be no longer any difficulty about that. It may be written down in the great speech.

I am not, in the first place, unaware that this Government has endured eighty-two years half slave and half free. I know that. I believe it has endured because during all that time, until the introduction of the Nebraska bill, the public mind did rest all the time in the belief that slavery was in course of ultimate extinction. I have always hated slavery, I think, as much as any Abolitionist — I have been an Old Line Whig — I have always hated it; but I have always been quiet about it until this new era of the introduction of the Nebraska bill began. I always believed that everybody was against it, and that it was in course of ultimate extinction.

The adoption of the Constitution and its attendant history led the people to believe so; and that such was the belief of the framers of the Constitution itself, why did those old men, about the time of the adoption of the Constitution, decree that slavery should not go into the new Territory, where it had not already gone? Why declare that within twenty years the African Slave Trade, by which slaves are supplied, might be cut off by Congress? Why were all these acts? What were they but a clear indication that the framers of the Constitution intended and expected the ultimate extinction of that institution? And now, when I say, as I said in my speech, that Judge Douglas has quoted from, when I say that I think the opponents of slavery will resist the farther

spread of it, and place it where the public mind shall rest with the belief that it is in course of ultimate extinction, I only mean to say that they will place it where the founders of this Government originally placed it.

I have said a hundred times, and I have now no inclination to take it back, that I believe there is no right, and ought to be no inclination, in the people of the Free States to enter into the Slave States, and inwith the question of slavery at all.

So much, then, for the inference that Judge Douglas draws, that I am in favor of setting the sections at war with one another. I know that I never meant any such thing, and I believe that no fair mind can infer any such thing from anything I have ever said.

Now, in relation to his inference that I am in favor of a general consolidation of all the local institutions of the various States. I have said, very many times, in Judge Douglas's hearing, that no man believed more than I in the principle of self-government; that it lies at the bottom of all my ideas of just government, from beginning to end. I think that I have said it in your hearing, that I believe each individual is naturally entitled to do as he pleases with himself and the fruit of this labor, so far as it in no wise interferes with the right of no other State, and that the General Government, upon principle, has no right to interfere with anything other than that general class of things that does concern the whole. I have said that at all times. I have said, as illustrations, that I do not believe in the right of Illinois to interfere with the cranberry laws of Indiana, the oyster laws of Virginia, or the liquor laws of Maine.

How is it, then, that Judge Douglas infers, because I hope to see slavery put where the public mind shall rest in belief that it is in the course of ultimate

extinction, that I am in favor of Illinois going over and interfering with the cranberry laws of Indiana? What can authorize him to draw any such inference? I suppose there might be one thing that at least enabled him to draw such an inference that would not be true with me or many others, that is, because he looks upon all this matter of slavery as an exceedingly little thing, — this matter of keeping one-sixth of the population of the whole nation in a state of oppression and tyranny unequaled in the world. He looks upon it as being an exceedingly little thing, — only equal to the question of the cranberry laws of Indiana; as something having no moral question in it; so little and so small a thing that he concludes, if I could desire that if anything should be done to bring about the ultimate extinction of that little thing, I must be in favor of bringing about an amalgamation of all the other little things in the Union. Now, it so happens — and there, I presume, is the foundation of this mistake — that the Judge thinks thus; and it so happens that there is a vast portion of the American people that do not look upon that matter as being this very little thing. They look upon it as a vast moral evil; they can prove it as such by the writing of those who gave us the blessings of liberty which we enjoy, and that they so looked upon it, and not as an evil merely confining itself to the States where it is situated; and... we agree that, by the Constitution we assented to, in the States where it is exists, we have no right to interfere with it, because it is in the Constitution; and we are by both duty and inclination to stick by that Constitution, in all its letter and spirit, from beginning to end.

Name: _____

Information Measure
Lincoln-Douglas Debates

Directions: Circle the letter of the best answer for each question below.

1. What does Lincoln say about the framers of the constitution?

   A. they prohibited slavery from the new territories
   B. they expected slavery to last
   C. they abolished free states
   D. they approved slavery in the southern states

2. What does Douglas claim that Lincoln will have to do to end slavery in all the States?

   A. preach a violent crusade against slavery in the South
   B. propose change in the Constitution
   C. pass the Kansas-Nebraska bill
   D. destroy the rights of the citizens

3. What does Lincoln say Douglas thinks is a "little thing"?

   A. the principle of self-government
   B. Northern pride
   C. oyster laws in Virginia
   D. slavery

4. What does Douglas believe about Congress?

   A. it doesn't want to interfere with States' rights
   B. it has no authority over states' decisions about slavery
   C. it has the Constitution's support to decide about slavery
   D. it will support free and slave states

5. What is Douglas saying the Free-Soil Party is actually doing with respect to the South and slavery?

   A. increasing the South's economic dependence on slavery
   B. discouraging slavery from spreading
   C. perpetuating slavery
   D. slowly ending slavery

6. What did Lincoln say in his Springfield speech?

   A. slavery will either spread to all states or end in all states
   B. the country will continue to be divided
   C. slavery should be extinct
   D. all states should have uniform policies

7. In which state did slavery disappear, according to Douglas?

   A. Kentucky
   B. New Jersey
   C. Illinois
   D. Virginia

8. Who does Lincoln quote as saying that "a house divided against itself cannot stand"?

   A. Shakespeare
   B. Stephen A. Douglas
   C. Thomas Jefferson
   D. himself

9. What is the major focus of Lincoln's speech?

   A. arguing for sectionalism
   B. justifying his plans for a civil war
   C. proposing changes in the Constitution
   D. denying Douglas' description of his speech

10. Douglas says "they stand on this side of the Ohio River and shoot across. They stand in Bloomington and shake their fists at the people of Lexington." What does he mean?

    A. people in South Carolina and Kentucky have different views on slavery
    B. Lincoln is advocating a war
    C. Lincoln will interfere non-violently with slave states
    D. abolitionists are getting ready to invade Kentucky

11. Why did Lincoln believe that slavery was permitted to endure for 82 years in the U.S.?

    A. everyone felt that slavery would eventually become extinct
    B. most people were willing to accept slavery
    C. the South depended on slavery
    D. the Constitution did not prohibit slavery

12. Why does Douglas refer to Austrians and Russians in relation to Lincoln's proposal?

    A. their constitutions were too easily changed
    B. they experienced civil war
    C. they tried to abolish slavery
    D. their governments were destructive to citizens' rights

13. According to Lincoln, what caused him to speak out against slavery?

    A. many people believed that slavery was a "vast moral evil"
    B. the framers of the Constitution opposed slavery
    C. a new era had begun with the introduction of the Nebraska bill
    D. he wanted to persuade the South against slavery

14. What is the best way to abolish slavery, according to Douglas?

    A. to have gradual emancipation
    B. to have all Free-Soil states band together
    C. to have a constitutional amendment
    D. to change the Kansas-Nebraska bill

Answer:

1 A. 2 B. 3 D. 4 B. 5 C. 6 A. 7 B. 8 D. 9 D. 10 C.

11 A. 12 D. 13 C. 14 A.

Name _____

Attitude Measure
(Reading Comprehension)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name: _____

WRITING ASSIGNMENT

Imagine that it is 1858 and you are an educated citizen living in Illinois. Because you are interested in politics and always keep yourself well-informed, you make a special trip to hear Abraham Lincoln and Stephen Douglas debating during their campaigns for the Senate seat representing Illinois. After the debates you return home, where your cousin asks you about some of the problems that are facing the nation at this time.

Write an essay in which you explain the most important ideas and issues your cousin should understand. Your essay should be based on two major sources: (1) the general concepts and specific facts you know about American history, and especially what you know about the history of the Civil War; (2) what you have learned from the readings yesterday.

Be sure to show the relationships among your ideas and facts.

Name _____

Attitude Measure
(Essay)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Thinking Questionnaire

Name: _____ Sex: _____

Date: _____

Directions: A number of statements which people have used to describe themselves are given below. Read each statement and indicate how you thought while writing the essay. Find the word or phrase which best describes how you thought and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which seems to describe how you thought during the task.

		Not at All	Somewhat	Moderately So	Very Much So
1.	It was important to do well on the task.	1	2	3	4
2.	I asked myself questions about the task.	1	2	3	4
3.	I reworded the task so I could understand it better.	1	2	3	4
4.	I was aware of my ongoing thinking processes.	1	2	3	4
5.	I tried to do my best on the task.	1	2	3	4
6.	I carefully planned my course of action.	1	2	3	4
7.	It was ok to guess on the task since it did not count.	1	2	3	4
8.	I had multiple thinking techniques or strategies to solve the task.	1	2	3	4
9.	I was aware of my own thinking.	1	2	3	4
10.	I thought through the meaning of the essay assignment before I began to write.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

	Not at All	Somewhat	Moderately So	Very Much So
11. I asked myself questions about what a task required me to do before I did it.	1	2	3	4
12. I judged the correctness of my work.	1	2	3	4
13. I evaluated the appropriateness of the thinking techniques or strategies that I used.	1	2	3	4
14. I was aware of which thinking technique or strategy to use and when to use it.	1	2	3	4
15. The task was challenging and interesting even if it didn't count for a school grade.	1	2	3	4
16. I tried to determine what the task was and what it required.	1	2	3	4
17. I knew how to recover from errors.	1	2	3	4
18. I selected and organized relevant information to solve the task.	1	2	3	4
19. I was aware of my trying to understand the task before I attempted to solve it.	1	2	3	4
20. I concentrated fully when doing the task.	1	2	3	4
21. I determined how to solve the task.	1	2	3	4
22. As I proceeded through the task, I asked myself, how well was I doing?	1	2	3	4
23. I kept track of my progress and, if necessary, I changed my techniques or strategies.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

	Not at All	Somewhat	Moderately So	Very Much So
24. When I solved the task I was aware of checking how well I was doing.	1	2	3	4
25. I worked hard on the task even if it did not count.	1	2	3	4
26. I set useful goals for myself.	1	2	3	4
27. I kept track of my progress.	1	2	3	4
28. The task was useful to check my knowledge.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

Self-Analysis Questionnaire

Directions: The next sentences are about your feelings or thoughts when you were taking the task. Read each statement and decide how you felt during the task. Find the word or phrase which best describes how you felt and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which best describes how you felt during the task.

	Not at All	Somewhat	Moderately So	Very Much So
29. I felt regretful.	1	2	3	4
30. I was afraid that I should have studied more for this test.	1	2	3	4
31. I felt that others would be disappointed in me.	1	2	3	4
32. I felt I may not have done as well on this test as I could.	1	2	3	4
33. I did not feel very confident about my performance on this test.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

Name_____
School_____
Teacher_____

STUDENT QUESTIONNAIRE

This two day testing period tried out some new ways to see what students know about history. Please give us your honest feelings on the questions below. We will keep your answers confidential and your teacher will not see them.

Circle the letter of the best answer for each question below.

1. Which word best describes the five periods you have spent on this project? Choose one.

   a. interesting
   b. tiring
   c. confusing
   d. boring

2. Which part of the testing did you like best?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

3. Which part did you like least?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

4. Overall, how was your performance on the tasks over these two days?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

5. On the first test with 20 short answer items, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

6. On the multiple choice test about the reading passages, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

7. How was your performance on semantic mapping?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

8. How was your performance on the essay?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

9. Which of these statements best describes how you wrote the essay?

   a. I was able to put together new ideas from the reading with other facts and ideas I already knew.
   b. I focused mostly on writing style
   c. I used yesterday's reading passages for most of my ideas
   d. I mostly used information I already knew before I read the passages.

10. Where did you learn most of the information you used on the 20-item short answer test?

    a. the teacher's lectures
    b. the textbook for this class
    c. other classes I've had
    d. televisions
    e. other reading outside of school
    f. I don't know

11. How well do you think these tests show what you know about this subject?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

12. How hard did you try on these tasks?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

13. Compared to how you usually try on tests your teacher gives, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

14. Compared to standardized tests you take once a year or so, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

15. Compared to homework,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

16. Compared to regular class discussions,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

17. The essay directions asked you to imagine you were in the same historical period as the authors of the reading passages and to explain the meaning of events to another person. How did that part of the directions influence your performance on the essay?

a. it helped me organize and choose information in my writing
b. it neither helped or hurt
c. it interfered with my writing
d. it made it fun
e. I didn't pay any attention to it at all

18. Which of these statements is true for you? Check all that apply

a. I didn't know the information the test asked for
b. The essay asked me to write in a way that I haven't been taught to do
c. These tests were more fun than regular tests
d. These tests were about the right level of difficulty

19. No matter how I did, I feel like I know the topic tested

a. very well
b. pretty well
c. only a little
d. not at all

20. How do you rate yourself as a writer?

a. I am a very good writer
b. I am a pretty good writer
c. I am a fair writer
d. I am not at all good at writing

21. How good do you think you are in history?

a. I am very good in history
b. I am pretty good in history
c. I am fair in history
d. I am not at all good in history

22. How good a student are you?

a. I am a very good student
b. I am a pretty good student
c. I am a fair student
d. I am not a good student

23. How do you think you could improve your performance on the essay task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

24. How do you think you could improve your performance on the semantic mapping task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

25. Which sequence did you follow?

a. read first two texts, made semantic map, then read additional texts
b. read first two texts and additional texts, then made semantic map
c. read first two texts, then made semantic map while exploring other texts
d. other (please specify) _____

_____ _____

26. Circle the appropriate number

How often do you:		At least once a week	A few times a week	once a month	less than once a month
A.	Take tests in history readings	1	2	3	4
B.	Write in-class essays in history	1	2	3	4
C.	Have homework in history	1	2	3	4
D.	Write in-class essays in other classes	1	2	3	4
E.	Write history essays at home	1	2	3	4
F.	Do longer research papers in history	1	2	3	4
G.	Take short answer tests in history	1	2	3	4
H.	Take multiple-choice tests in history	1	2	3	4

Appendix B

Content Assessment Materials: Chinese Immigration

UCLA Content Assessment Project
(Field Testing, July, 1991)

Student Demographic Form

Student's Name _____

Student's ID _____

Birthday _____

Teacher's Name _____

School in 1990-1991 _____

Grade Level in Sept. 1991 _____

Please circle the appropriate number:

Gender:
- Male 1
- Female 2

Ethnicity:
- White 1
- Black 2
- Hispanic 3
- American Indian 4
- Asian/Pacific Islander 5
- Other 8

Name: _____

How Much Do You Know About U.S. History?

Directions: This is a list of terms related to U.S. History. Many of them are related to 19th century Asian immigration, but some of them are from other periods in U.S. History.

In the space after each term, write down what comes to mind when you think of that term in the context of U.S. History. A brief definition would be acceptable, or a brief explanation of why that person, place, or thing was important. If the term is general, such as "Civil rights," give both a general definition and a specific example of how the term fits into U.S. History, if you can.

Good Example: CIVIL RIGHTS. Rights guaranteed to all citizens regardless of *race, sex, religion, etc. Blacks fought for their civil rights in the 1960s. Martin Luther King, Montgomery bus boycott.*

Do not define the term by simply restating the same words.

Bad Example: SURVIVAL OF THE FITTEST -- only the fittest survive.

Even if you are not sure about your answer, but think you know something, feel free to guess.

There are probably more items here than you will be able to answer in the time given. Start with the ones you know best, and work quickly so that you can answer as many as possible. Then go back and answer the ones of which you are less sure. Do not spend too much time on one specific item.

1. Progressive Era

_____

_____

2. Japanese internment camps

_____

_____

3. Immigration Acts of 1920s

_____

_____

4. popular sovereignty

5. Gold Rush

6. working class

7. Americanization

8. Yellow Peril

9. Promontory Point, Utah

10. Gentlemen's Agreement

11. nativism

12. Boston Tea Party

13. industrialization

14. Cold War

15. ethnicity

16. Manifest Destiny

17. naturalization

18. Chinese Exclusion Act of 1882

19. "city upon the hill"

20. Unalienable Rights

Name _____

Attitude Measure
(Prior knowledge)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name:_____

BROOKS-PIXLEY

Debate on Chinese Immigration*

1877

In the late 1800's and early 1900's there were different opinions and strong feelings about the immigration of various groups of people to the United States. The following texts concerning Chinese immigration are taken from a debate held in 1876 between Frank Pixley, the attorney representing the city of San Francisco, and B. S. Brooks, the attorney on behalf of the Chinese. Keep in mind that the texts were taken from a period in history when there was great apprehension and opposition to certain ethnic groups such as those immigrating from Asia. Be aware that the issues and language, though condensed, have not been altered in any way from the original debate.

FRANK M.PIXLEY

If I understand the scope of this investigation, a joint committee of the two houses of Congress is here for the purpose, as it were, of taking testimony, and reporting their conclusions to the Congress of the United States, in reference to the propriety of encouraging or restraining Chinese immigration from the Asiatic empire to this coast, and to this port of San Francisco. If I understand our position, we are here like as in a court, and we are required on the part of those who seek to oppose Chinese immigration to make a statement of our case, and what expect to prove in that particular.

"For what purpose do they come? With the intention of remaining and making the United States their home, or returning to China when they have acquired a competence?" Our answer to "for what purpose do they come" is

* From *Report of the Joint Special Committee to Investigate Chinese Immigration* (Washington, Government Printing Office, 1877).

embraced in the single word "money." They come for coin. They come from poverty, from destitution, from low wages, from bad government, from a redundant and overwhelming population to a free government, to liberty, protection, labor, remunerative wages, and the object of their coming is that they may obtain here by their wages money enough to return and enjoy their accumulation in the land of their birth. It is in testimony, or will be, that two or three or four hundred dollars is a competency; five hundred dollars is independence; a thousand dollars is a liberal fortune. Upon it they may exist, because the cost of living is confined to but a few pence or a few cents a day.

"What kind of labor do they perform" Mr. King will dilate upon this at greater length than I shall. I will only say that they perform all kinds of light labor, and that particularly which requires no capital; and they are expert in that which requires dextrous manipulations of the fingers—as the assorting of wool, working in silks, the rolling of cigars, and such matters as that. They are imitative and quick to learn and they have monopolized many of the branches of our industry. Laundry-work, cigar-making, slippers, sewing-machine labor, they have nearly monopolized. They are largely employed as domestic servants and as office-boys. In assorting and repacking teas, in silk and woolen manufactories, in fruit-picking, in gardening, in harvesting, in building levees for the restoration of tule lands, in railroad-building, in placer-mining, in basket-peddling of vegetables and fruits, in fishing and peddling fish, are among the most noted of their industries, and from these industries that I have named they have nearly driven out the entire white labor. They do not, as a rule, work in underground mines, nor in tunnels, nor in heavy stone-work. They are rarely found in the forest; they are rarely used as teamsters, for heavy hauling; and, as a rule, they never perform any work that is both

heavy and dangerous, or that is heavy or dangerous. To say what they do, one of the strongest points is what they do not do. They have introduced into our State not one single one of the peculiar industries of China. In our earlier and gushing period over the Chinese, we said to ourselves: "They will introduce here the culture of tea and rice, and the manufacture of silks; we shall have all their curious industries, and all their new productions." Not one acre of land has yet been devoted to the culture of rice; not one shrub to the production of tea; not one single industry has been introduced, so far as I am advised, that is peculiar to the Chinese people.

"How does their employment affect white labor?" and here comes the questions, "How does their employment affect our white labor?" We answer, and this is the burden of our arraignment of this Chinese immigration; it is not our sympathy for the wealthy classes, it is our sympathy for the labor classes, upon which, whatever may be our feelings, depends really the whole superstructure of our Government. The true American hero is the man who takes his dinner out in his tin plate, works all day, six days in the week, and brings his wages home for his wife to expend in the maintenance and education of the family, in their clothing and in their protection. Chinese labor drives this class of people from the field. It drives them to starvation. It is a competition that they cannot undertake. The white race, owing to centuries of physical treatment, is incompetent to enter upon the race. The man who labors in our streets and city, and in our country, has been, as have his fathers for generations before him, fed on meat and bread. He demands meat and bread to maintain his physical strength and his existence. Meat and bread command more money than labor will pay for at Chinese rates in any place in America, and especially in California. The Chinaman from generations has

been in the habit of living upon rice, tea, dried fish, and desiccated vegetables. The kind of food which will support the Chinaman can in San Franscisco be purchased for ten cents a day, and the kind of food which is required to support an American or European laborer cannot be bought for several times that amount. The American laborer has other matters that he may not set aside. He has a wife. The Chinaman is an adult male who has no wife, no family, no child. Our white laborers are, as a rule, married, and fathers and heads of families, and according to our mode of civilization the poorest laborer with the poorest wife must occupy a room by himself for his bed and must have at least another room to cook and eat in. If he has a boy and a girl growing to ages of puberty, the boy must have a room for himself and the girl must have a room for herself, and both must be separate from the parents' bed. It is the ingrained decency of our civilization. It is as impossible to change it as it is to change us from the worship of the Christian God to the heathen tablet. Chinamen in a double room like this would throw a partition through the center and build bunks on the side and lie down upon the floor. They would cook their tea and dried rice in a brazier not bigger than a spittoon. One hundred of them would live in this room, while the poorest Christian family of five in the State would think themselves crammed in double the space. In your minds you may drift off to see how this affects rents.

We admit that Chinese labor has contributed to the more speedy development of our material resources. We acknowledge the advantage it has been to certain industries, and that many individuals have become richer than they would expect for the presence of the Chinese. We admit their convenience to us as domestic servants. We do not represent the Chinese as wanting in many of the essentials of good citizens. The burden of our accusation against

them is that they come in conflict with our labor interests; that they can never assimilate with us; that they are a perpetual, unchanging, and unchangeable alien element that can never become homogeneous; that their civilization is demoralizing and degrading to our people; that they degrade and dishonor labor; that they can never become citizens, and that an alien, degraded labor class, without desire of citizenship, without education, and without interest in the country it inhabits, is an element both demoralizing and dangerous to the community within which it exists.

B. S. BROOKS

I do not sympathize at all with the view of the subject which has been presented on the other side. The very people who raise all this clamor, who fill the halls, pass resolutions and elect delegates, would never have been in this country, if their views had prevailed. It seems strange to me that one class of emigrants should be permitted to rise against another class of emigrants, because they come in competition with *them*. I deny the right of any foreigner, who comes to this country, to do that. We permit them to come here. They come here by virtue of our laws. No foreigner has the right *as a foreigner*, simply to come into this country, and to establish himself and become an owner of the soil. It is *our law* which gives him that right. I have no sort of sympathy for the argument made by an Irishman, a German, or a foreigner of any nation, who has come here and been naturalized, and been made a citizen, and allowed to hold land, when he talks about *our* land being land for the *white* man, and says that this yellow colored man comes in competition with the white man. It is nothing to me if he does. I do not think it concerns the nation or humanity, or the world at large, that the yellow man's labor comes in competition with the black man's, the red man's, or the labor of any other man. I do not subscribe to the creed of my friend on the other side. I believe these men *have* souls. I believe in the common humanity and brotherhood of all men. I do not claim any rights whatever as against a red man, or a black man, or a yellow man. If he can compete with me on a fair footing, let him compete. If he diminishes my earnings, I have no right to complain. He has as good a right to earn a living on God's footstool as I.

When a stream of water overflows its banks and becomes a flood, it is a terrible engine of destruction, but when it runs its natural course and is used and utilized, what can be more beneficent? The Chinese element is an element of prosperity, of future greatness, of wealth, but you can make an evil of it, as you can of anything. When you look at this matter, I think you will see that all that is noxious about it comes from ourselves, and not from the Chinese.

Until the land can be profitably used for the cultivation of grain, and until the grain can be got out to market, it is not open to the settlement of white families. As *pasture land*, this land will support about one white man to the thousand acres, and the herdsman who follows the cattle, has no family, and is a simple Bedouin. As *agricultural land*, this same land would support a *hundred* people to the 1,000 acres, and these not nomads, but families in homesteads, with villages, schools, and temples of the living God, whom the Chinamen and the Americans, some of them, worship. The effect of the railroads which the Chinese have built, is to convert these valleys from simple pasture lands into farm lands; to open them for white people.

There are about 5,000,000 acres under cultivation, as near as I can calculate it, in the State. There are about 5,000,000 acres more in a natural state which can be used in the same way. Now, the Chinamen have not only made homes and furnished employment for white men, but they have given a living, the bread and butter to 500,000 white men. That is the effect of the hundred thousand Chinamen here. I do not stand here to plead for these hundred thousand Chinamen. I stand here to plead for the interests of these 500,000 white men and women and children, and I will plead for them against their own wishes, if necessary.

I asked a former Surveyor General of this State to estimate the increase in the value of the property of this State created by Chinese labor in building railroads, and in reclaiming tule lands alone, and the amount he gave me is $289,700,000. That is the wealth which a hundred thousand Chinamen had added to California. It is wealth owned, held and enjoyed by white men and not by Chinamen. The Chinamen do not carry it away with them; they could not, even if they wished to do so.

Many men have spoken to me, wishing me God speed, and said they must desert the country, without Chinese labor.

These men will come before you; I shall bring witnesses before you from all parts of the State, who stand high in the opinion of their fellow-citizens, and they will tell you that they cannot do without this labor, and that the State cannot do without it. It is not a question whether we are benefitting the Chinese, and whether they want to stay or go. If I had my will about it I would fence them in and not let them go, but that I acknowledge their right to go where they please. They pick the fruit and they pack the fruit. This is going to be one of the great interests of this State, and one that we must rely upon, and which I am glad to see is developing splendidly.

I shall show by testimony before you that the Chinese have added vastly to the wealth of the State, that they have opened homes for half a million of white people—five times their own number—and that there is ample scope in the future for them to go on doing the same thing; that we have only utilized some five millions acres of this vast domain, and that there are some 50,000,000 acres more which we can yet utilize, but we must introduce new industries.

I do not think the Chinamen are all angels, and I do not think everything about them is lovely. There are a great many things about every foreign people which I dislike. I have my American prejudices, but that would not cause me to do them injustice.

It is said that this is a terrible criminal element, but what is the state of our city? Here is a city of 250,000 inhabitants, a big city, and a commercial city. Every one knows that in a port vice congregates; and yet this whole city is kept in order by only 150 policemen.

After having heard some of the accounts of this terrible cesspool, you would think of course it requires at least 500 policemen to take care of it; but this whole city has but 150 policemen. Yet, they make more arrests than any other 150 policemen probably in the world. I say, and I do not believe anybody will deny it who will inquire into the matter, that there is no city in the world of its size where property and life are as safe as in San Francisco. I think the statistics will show it. If you take the police report you will find that the arrests in the year are 20,180. That is pretty good for 150 policemen. Out of that number of arrests 7,643 are for drunkenness. More than one third are for drunkenness and not one of these is a Chinaman. I do not think any man ever saw a drunken Chinaman on our streets. I do not myself remember ever to have seen one during my entire residence in this country.

I will not detain the Committee longer than to say that on all these points we shall call witnesses, and the Committee will find that they are among the best people in the State of California, and those most competent to know the merits of this question.

Name _____

Information Measure
Pixley - Brooks

Directions: Circle the letter of the best answer for each question below.

1. According to Brooks, what would it mean to utilize California's land fully?

   a. convert pasture lands to farm lands
   b. allow more immigrants to own land
   c. use more efficient farming methods
   d. increase immigration

2. How does Pixley characterize Chinese workers?

   a. polite
   b. imitative
   c. strong
   d. inventive

3. According to Pixley, why could Chinese laborers work for less money?

   a. they brought money from China
   b. they did not have to support families
   c. they did not spend money on recreation
   d. they lived on farms

4. Which industry does Pixley say the Chinese have nearly monopolized?

   a. cigar making
   b. silk production
   c. underground mining
   d. tea growing

5. What does Pixley claim about the Chinese?

   a. they assimilate easily to American culture
   b. they have not contributed to American industrialization
   c. they make valuable citizens
   d. their civilization is demoralizing

6. What does Brooks say was the result of building railroads in California?

   a. decrease in agricultural land
   b. increase in pollution
   c. decrease in transportation costs
   d. increase in housing for whites

7. Why does Pixley believe that the "true American hero" cannot compete with the Chinese worker?

   a. the Chinese are industrious and patient workers
   b. the Chinese are more polite and conscientious
   c. the Chinese need less money to live on
   d. the Chinese have no political duties and do not pay taxes

8. For which of the following does Brooks argue?

   a. using Chinese labor to develop land
   b. passing laws to prevent Chinese immigration
   c. allowing Chinese immigrants to become citizens
   d. forming a labor union for the Chinese

9. What is Brooks' main argument?

   a. California has more white people than Chinese people
   b. Chinese immigrants contribute to the wealth of California
   c. Chinese immigrants have a low crime rate
   d. Chinese people are good citizens

10. According to Brooks, what was the most important contribution of Chinese immigrants?

    a. increasing the labor force
    b. introducing new farming methods
    c. harvesting produce
    d. opening new lands

11. In Pixley's view, why do people come to the United States from China?

    a. to escape religious persecution
    b. to become "good Americans"
    c. to make money
    d. to learn a trade

12. Why does Brooks deny the right of one group of immigrants to rise against another group of immigrants?

    a. all immigrants contribute to the United States
    b. California needs all immigrants
    c. all immigrants have the same rights
    d. all immigrants eventually become naturalized

Name _____

Attitude Measure
(Reading Comprehension)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name _____

WRITING ASSIGNMENT

Imagine that it is 1876 and you are an educated citizen living in California. Because you are interested in immigration, you make a special trip to hear the debate about Chinese immigration between Frank Pixley, the attorney representing the city of San Francisco, and B. S. Brooks, the attorney on behalf of the Chinese. When you return home, your cousin asks you about some of the concerns about Asian immigration in the country.

Write an essay in which you explain the most important ideas and issues your cousin should understand. Your essay should be based on two major sources: (1) the general concepts and specific facts you know about American History, and especially what you know about Asian immigration to this country; and (2) what you have learned from the readings.

Be sure to show the relationships among your ideas and facts.

Name _____

Attitude Measure
(Essay)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Thinking Questionnaire

Name: _____ Sex: _____

Date: _____

Directions: A number of statements which people have used to describe themselves are given below. Read each statement and indicate how you thought while writing the essay. Find the word or phrase which best describes how you thought and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which seems to describe how you thought during the task.

		Not at All	Somewhat	Moderately So	Very Much So
1.	It was important to do well on the task.	1	2	3	4
2.	I asked myself questions about the task.	1	2	3	4
3.	I reworded the task so I could understand it better.	1	2	3	4
4.	I was aware of my ongoing thinking processes.	1	2	3	4
5.	I tried to do my best on the task.	1	2	3	4
6.	I carefully planned my course of action.	1	2	3	4
7.	It was ok to guess on the task since it did not count.	1	2	3	4
8.	I had multiple thinking techniques or strategies to solve the task.	1	2	3	4
9.	I was aware of my own thinking.	1	2	3	4
10.	I thought through the meaning of the essay assignment before I began to write.	1	2	3	4
11.	I asked myself questions about what a task required me to do before I did it.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

	Not at All	Somewhat	Moderately So	Very Much So
12. I judged the correctness of my work.	1	2	3	4
13. I evaluated the appropriateness of the thinking techniques or strategies that I used.	1	2	3	4
14. I was aware of which thinking technique or strategy to use and when to use it.	1	2	3	4
15. The task was challenging and interesting even if it didn't count for a school grade.	1	2	3	4
16. I tried to determine what the task was and what it required.	1	2	3	4
17. I knew how to recover from errors.	1	2	3	4
18. I selected and organized relevant information to solve the task.	1	2	3	4
19. I was aware of my trying to understand the task before I attempted to solve it.	1	2	3	4
20. I concentrated fully when doing the task.	1	2	3	4
21. I determined how to solve the task.	1	2	3	4
22. As I proceeded through the task, I asked myself, how well was I doing?	1	2	3	4
23. I kept track of my progress and, if necessary, I changed my techniques or strategies.	1	2	3	4
24. When I solved the task I was aware of checking how well I was doing.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

	Not at All	Somewhat	Moderately So	Very Much So
25. I worked hard on the task even if it did not count.	1	2	3	4
26. I set useful goals for myself.	1	2	3	4
27. I kept track of my progress.	1	2	3	4
28. The task was useful to check my knowledge.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

Self-Analysis Questionnaire

Directions: The next sentences are about your feelings or thoughts when you were taking the task. Read each statement and decide how you felt during the task. Find the word or phrase which best describes how you felt and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which best describes how you felt during the task.

	Not at All	Somewhat	Moderately So	Very Much So
29. I felt regretful.	1	2	3	4
30. I was afraid that I should have studied more for this test.	1	2	3	4
31. I felt that others would be disappointed in me.	1	2	3	4
32. I felt I may not have done as well on this test as I could.	1	2	3	4
33. I did not feel very confident about my performance on this test.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

Name_____
School_____
Teacher_____

STUDENT QUESTIONNAIRE

This two day testing period tried out some new ways to see what students know about history. Please give us your honest feelings on the questions below. We will keep your answers confidential and your teacher will not see them.

Circle the letter of the best answer for each question below.

1. Which word best describes the five periods you have spent on this project? Choose one.

   a. interesting
   b. tiring
   c. confusing
   d. boring

2. Which part of the testing did you like best?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

3. Which part did you like least?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

4. Overall, how was your performance on the tasks over these two days?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

5. On the first test with 20 short answer items, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

6. On the multiple choice test about the reading passages, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

7. How was your performance on semantic mapping?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

8. How was your performance on the essay?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

9. Which of these statements best describes how you wrote the essay?

   a. I was able to put together new ideas from the reading with other facts and ideas I already knew.
   b. I focused mostly on writing style
   c. I used yesterday's reading passages for most of my ideas
   d. I mostly used information I already knew before I read the passages.

10. Where did you learn most of the information you used on the 20-item short answer test?

    a. the teacher's lectures
    b. the textbook for this class
    c. other classes I've had
    d. televisions
    e. other reading outside of school
    f. I don't know

11. How well do you think these tests show what you know about this subject?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

12. How hard did you try on these tasks?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

13. Compared to how you usually try on tests your teacher gives, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

14. Compared to standardized tests you take once a year or so, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

15. Compared to homework,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

16. Compared to regular class discussions,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

17. The essay directions asked you to imagine you were in the same historical period as the authors of the reading passages and to explain the meaning of events to another person. How did that part of the directions influence your performance on the essay?

a. it helped me organize and choose information in my writing
b. it neither helped or hurt
c. it interfered with my writing
d. it made it fun
e. I didn't pay any attention to it at all

18. Which of these statements is true for you? Check all that apply

a. I didn't know the information the test asked for
b. The essay asked me to write in a way that I haven't been taught to do
c. These tests were more fun than regular tests
d. These tests were about the right level of difficulty

19. No matter how I did, I feel like I know the topic tested

a. very well
b. pretty well
c. only a little
d. not at all

20. How do you rate yourself as a writer?

a. I am a very good writer
b. I am a pretty good writer
c. I am a fair writer
d. I am not at all good at writing

21. How good do you think you are in history?

a. I am very good in history
b. I am pretty good in history
c. I am fair in history
d. I am not at all good in history

22. How good a student are you?

a. I am a very good student
b. I am a pretty good student
c. I am a fair student
d. I am not a good student

23. How do you think you could improve your performance on the essay task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

24. How do you think you could improve your performance on the semantic mapping task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

25. Which sequence did you follow?

a. read first two texts, made semantic map, then read additional texts
b. read first two texts and additional texts, then made semantic map
c. read first two texts, then made semantic map while exploring other texts
d. other (please specify) _____

_____

26. Circle the appropriate number

How often do you:	At least once a week	A few times a week	once a month	less than once a month
A. Take tests in history readings	1	2	3	4
B. Write in-class essays in history	1	2	3	4
C. Have homework in history	1	2	3	4
D. Write in-class essays in other classes	1	2	3	4
E. Write history essays at home	1	2	3	4
F. Do longer research papers in history	1	2	3	4
G. Take short answer tests in history	1	2	3	4
H. Take multiple-choice tests in history	1	2	3	4

Appendix C

Content Assessment Materials: General Immigration

UCLA Content Assessment Project
(Field Testing, July, 1991)

Student Demographic Form

Student's Name _____

Student's ID _____

Birthday _____

Teacher's Name _____

School in 1990-1991 _____

Grade Level in Sept. 1991 _____

Please circle the appropriate number:

Gender:
- Male 1
- Female 2

Ethnicity:
- White 1
- Black 2
- Hispanic 3
- American Indian 4
- Asian/Pacific Islander 5
- Other 8

Name: _____

How Much Do You Know About U.S. History?

Directions: This is a list of terms related to U.S. History. Many of them are related to 20th century immigration, but some of them are from other periods in U.S. History.

In the space after each term, write down what comes to mind when you think of that term in the context of U.S. History. A brief definition would be acceptable, or a brief explanation of why that person, place, or thing was important. If the term is general, such as "Civil rights," give both a general definition and a specific example of how the term fits into U.S. History, if you can.

Good Example: CIVIL RIGHTS. Rights guaranteed to all citizens regardless of *race, sex, religion, etc. Blacks fought for their civil rights in the 1960s. martin Luther King, Montgomery bus boycott.*

Do not define the term by simply restating the same words.

Bad Example: SURVIVAL OF THE FITTEST -- only the fittest survive.

Even if you are not sure about your answer, but think you know something, feel free to guess.

There are probably more items here than you will be able to answer in the time given. Start with the ones you know best, and work quickly so that you can answer as many as possible. Then go back and answer the ones of which you are less sure. Do not spend too much time on one specific item.

1. 14th Amendment

_____

_____

2. immigration quotas

_____

_____

3. Brown vs. Board of Education

_____

_____

4. Constitutionality

5. Immigration Acts of 1920

6. assimilation

7. Mexican War of 1848

8. restrictionists

9. American Federation of Labor

10. westward movement

11. nativism

12. New Deal

13. affirmative action

14. Americanization

15. ethnicity

16. Manifest Destiny

17. naturalization

18. Immigration Act of 1965

19. illegal alien

20. private enterprise

Name _____

Attitude Measure
(Prior knowledge)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name:_____

SIMON-GRAHAM

Debate on Immigration*

1981

There are different opinions about the immigration to the United States. The following texts are testimonies from a hearing concerning immigration held in 1981 by Julian Simon, a professor of sociology, and Otis Graham, Jr., a professor of history. The issues and language, though condensed, have not been altered in any way from the original debate.

JULIAN SIMON

Mr. Chairman, members of the subcommittee, I very much appreciate the invitation to give the subcommittee my views on the economic impact of immigrants upon natives. I will first summarize the study I did for the Selection Commission. Then I will make a few more general remarks about the impact of immigrants.

My general conclusion is that the average immigrant family takes less in welfare and pays more in taxes than the average native family. There are other impacts through the labor market, and effects on productivity add up to a large positive effort on the standard of living for natives.

In 1976, the survey of economic opportunity which was carried out by the Bureau of the Census gathered detailed data on the incoming use of social services by over 150,000 families, including about 15,000 immigrant families. The results show—this is from my study—that from the time of entry until about 12 years later, immigrants used substantially less of such public services as welfare and unemployment compensation payments, food stamps,

* From *Hispanic Immigration and Select Commission on Immigration's Final Report* (Washington, Government Printing Office, 1981).

medicare, medicaid, and schooling for children than do native families. This is largely due to less use of social security because of their youthful age. When they come, immigrants are young and strong.

Later, when the immigrant family retires and collects social security, it typically has raised children who are contributing taxes to social security, thereby balancing out the parents' receipts, just as the native families.

In this way, there is a one-time benefit to natives because the immigrants arrive without a generation of elderly parents who receive social security.

After about 3 to 6 years in the country, the average immigrant family comes to earn as much as the average native family and thereby pay as much in taxes as do native families.

So the net balance of these two forces, the taxes paid by the immigrants and the services received by immigrants, is positive in every year for natives; that is, immigrants contribute more to the public coffers than they take from them.

When you look at immigrants the way you look at investments, in such a social capital as dams or roads, immigrant families are an excellent investment, worth somewhere between $15,000 and $20,000, even calculated at relatively high rates for the social cost of capital. That was in 1975 dollars, and you can compare that $15,000 to $20,000 benefit to the, say, average $11,000 mean yearly earnings for an average native family in that year.

We are very properly worried about the social security system because the ratio of retired persons to those in the working age is becoming a difficult burden. Immigrants typically are additional young workers who help support our retired persons and bring no retired persons into the social security system

to pension. Hence, each immigrant family makes an immediate and large contribution toward reducing the social security burden.

Let us consider what is likely the most important long-run effect of immigrants: The impact on productivity of these additional workers and consumers is likely to dwarf all else after a few years in the country.

Some productivity increase arises from immigrants working in industries and laboratories in the United States that are at the forefront of world technique.

We benefit along with others from the contribution to world productivity in, say, genetic engineering that immigrants would not be able to make in their home countries.

Other increases in productivity come from increased production in particular industries through learning-by-doing and other gains from larger industry scale.

In sum, immigrants benefit natives through the public coffers by using less than their share of services and paying more than their share of taxes. They cover the additional public capital needed on their account through the debt service on past investments.

The same general welfare argument applies as for free trade, but it is cold comfort to the dislocated persons. Other than this inequity, immigrants viewed in economic terms seem an excellent bargain.

There are still other considerations that are harder to pin down with data but that are likely to bode well for natives. Immigrants are likely to save more, because they start out with less assets. They are likely to work harder because, not having assets and economic security, they are likely to act hungry.

For the same reason, they are likely to be more mobile in search of economic opportunity, and such mobility is crucial in keeping the economy in adjustment.

Immigrants seem to be more innovative than people who have never changed countries, and innovation is vital in boosting productivity.

And perhaps most important, immigrants—and especially the young people—are more hopeful and have more of a forward-looking outlook, and in a time when our Nation seems afflicted by paralysis, this hopeful economic view of the future of America as their land of opportunity must be important to the economy.

Why are we so worried about what we think is an additional burden now? Especially when the immigrants don't increase our burden but, rather, lighten it.

Back in the fifties we managed to give college educations to unprecedented numbers of ex-GI's, and at the same time start a vast costly Interstate Highway System—without a great sense of burden. We spent large proportions of our income for defense—again without feeling that we couldn't cope, though the burdens really were large.

What has happened to our spirit? And what has happened to our minds, when we can see that the one major problem facing us—supporting the aged—is directly helped by more immigrants?

Many of us have charity in our minds when we thing of immigration. This shouldn't cloud the fact that we are also bringing in immigrants for our own sake.

This is one of this rare opportunities where we can do well while we are doing good.

In that vein I will read some excerpts from a letter I received last year about an article I wrote that was reprinted in the Columbia, S.C. newspaper:

> DEAR PROFESSOR SIMON: Your article about the economic impact of immigrants reprinted here in The State Record in Columbia has encouraged me a lot. Your argument has really relieved my doubt about being a burden for this country. I am a Vietnamese refugee arriving in 1976.
>
> If the American public has a supporting and fair attitude...the immigrants would be able to be more productive. A negative and false opinion about our situation really discourage the young generation to stand up with dignity and to contribute to build up our society....Somehow the American public expects to hear another kind of message such as: truly the immigrants are burdensome, but the great country of U.S.A. will make sacrifices to help out.
>
> It is unhealthy. I think all immigrants should renounce to the idea of charity. Nobody can maximize the potential without being recognized and expected to do so.
>
> Signed, Chris Le

Americans were [once] quite sure that immigrants were good for the country. And it is still true that more immigrants are good for us, though they may impost costs in the very short run, and though some must suffer more adjustment costs than others. And we should let them in for our sake, not just for theirs, just as Mr. Le wrote.

OTIS L. GRAHAM, JR.

I am glad of the chance to speak to you as a professional historian who has reflected a great deal about how history might guide policymakers in all areas of their responsibility, even in the especially complex area of immigration and population.

Though I am enthusiastic about this committee's receptivity to the historical perspective, I must begin with a note of warning. Turning to history for guidance is very likely to mislead you. That is not the fault of history, but of the way it is too frequently seen and used.

It is difficult to imagine a decision that any of us takes which is not shaped by certain assumptions about the past, but we normally take a very simplistic view of what history teaches, and lead ourselves into error.

For it will be said—had already been said, by people making quick raid upon the past to confirm some bias of their own—that U.S. history shows that large-scale immigration in the years from the 1890's to World War I stirred up an unnecessary fuss—that the surge of immigration that so alarmed contemporaries did not harm the United States in ways that contemporaries feared, and that a strict limit was advocated by people who had racist and nativist outlooks.

In a general way, this does seem to be what our history reveals.

Many people then take the next and unwarranted step, concluding that history teaches that we should relax about today's surge of immigration, which since the mid-1960's has run at levels comparable to or exceeding those prior to World War I, and that those who raise alarm today and call for strict limits are moved by base motives and perceive problems where none exist.

But history does not teach by analogy. Circumstances change, and good historical analysis pays close attention to those changes.

Ernest May's book "Lessons of the Past" is a rich record of decisions taken by public leaders who remembered the past but who were very poor historical analysts—who always seemed to assume that because a thing had turned out badly, it should always be shunned, or because a thing had turned out well, it should always be repeated.

This is history by simplistic analogy. We get too much of it in the immigration debates of today.

Let me attempt a much too sketchy illustration. Two very wide-spread contemporary assumptions, probably valid down our historical experience, are now, it seems to me, untenable; things have changed.

The first of these assumptions is that the influx of additional population is beneficial to American society. Though not everyone comes to this realization at the same speed, it is increasingly clear that this is not so.

I acknowledge that this is a vast and complex subject, but an alert legislator, who has followed the studies and report of the 1972 population commission, of the recent Global 2000 report which derived from an impressive interagency effort, not to speak of the thousands of more specialized studies of the population dimension in the American future, of which the best summary is Gerald O. Barney, editorial, "The Unfinished Agenda"—that legislator who has a speaking acquaintance with these studies will know that the continent that absorbed the waves of immigration prior to World War I is now packed with a more numerous and ecologically destructive human population which even without immigration will grow for another 50 years.

A good case can be made that there are too many Americans already, for the margins of environmental safety and the standards of life which we wish to secure.

Certainly President Nixon's commission concluded this in 1972, and there are more of us today than then.

Perhaps in Washington there is the view that the only shortages are dollars and apartments, but any citizen knows that this Nation is pressing against the edges of many of our resources and our supporting environment.

Newsweek, on February 23, 1981, for example, ran a long essay on "The Browning of America," charting the water shortage that afflicts much of the Nation from west to east.

Energy is short, water is short, housing and land are short—but we must break into such language, as did Garrett Hardin, and point out that we have not so much a resource shortage as a people longage.

Time has fundamentally altered our circumstances. Immigration is now at the center of the question of American population size, matching the influence of domestic increase. This has not been historically the case in modern times.

If the alarms rung about large-scale immigration prior to World War I were largely exaggerated, that does not mean that they are to be so regarded today.

The second assumption which our history implants in the contemporary mind is equally false. That is: those who oppose large-scale immigration, who raise the immigration issue as a subject for "viewing with alarm," must be today, as they were in the distant past, bigots and nativists who did not wish to

make room for people different from themselves. But here, too, things have fundamentally changed.

The United States is a more ethnically diverse Nation today, has profoundly altered its basic racial and ethnic attitudes toward a more tolerant, multicultural pattern. Racism and nativism there may be, but they are on the defensive in our culture; "minorities" are not defenseless victims, not under attack from the dominant culture as once they were.

We could assume, back down our past, that immigration restrictionists were nativists, perhaps also members of the Ku Klux Klan. We cannot do so today. Restrictionists are moved by considerations of economic and ecosystem viability, by love of the American environment, by obligation to the unborn—not by dislike of aliens.

We cannot dismiss their arguments on the old grounds, for they are not arising from the old grounds.

Thus, I would suggest that our history on the immigration issue is a poor guide for contemporary policymakers. The past is often misleading, if we simplistically assume that it repeats itself.

But is there no more that the historian can tell you, beyond that some things have changed, and that the most prominent of these—the population/resource/environmental relationship and the source and nature of contemporary restrictionism—have transformed the issue and require us to look at it in a new way?

When you make immigration policy, you are making population policy for the United States, shaping the civil rights movement, influencing the economic structure.

Would you prescribe for such a society, in such a world, a population policy permitting the entrance of 800,000 annual legal migrants [as present policy permits] with an open end for additional refugee admissions, along with an unofficial policy of such porous borders that literally any number can enter and routinely as many enter around the official channels as through them?

You would certainly not, as legislators, approve a policy that would lead to such drastic population consequences, estimated by Leon Bouvier at between 100 and 300 million more Americans in 100 years than we would have in the absence of immigration.

But that is the immigration policy we have today, thus the population policy we have today. That is the larger context which you must have firmly in mind.

Name _____

Information Measure
Simon - Graham Passages

Directions: Circle the letter of the best answer for each question below.

1) What does Simon say is true of the average immigrant family after it has been in this country three to six years?

   a. it earns as much as the average native family
   b. it pays lower taxes than the average native family
   c. it pays higher taxes than the average native family
   d. it has finally come off welfare

2) What does Graham suggest about the massive influx of foreigners prior to the First World War?

   a. it was a terrible blow to the American economy
   b. it verified the fears of those who opposed open immigration policy
   c. it was not as bad for the country as many believed
   d. it created new industry

3) What does Simon's analysis suggest about the average immigrant family?

   a. it requires $15,000 in welfare support
   b. it pays $15,000 in taxes
   c. it takes less in services than it pays in taxes
   d. it takes more in services than it pays in taxes

4) What does Simon believe about immigrants?

   a. they burden the economy
   b. they bring mobility to the job market
   c. they pay an unfair amount of taxes
   d. they are not as innovative than other workers

5) Which of the following motivates present-day immigration restrictionists, according to Graham?

   a. racism
   b. fear of alien cultures
   c. belief in ethnic purity
   d. love of the environment

6) What is the basis of Simon's argument?

a. the nation needs cheap labor
b. workers need to be shifted into new industries
c. immigrants bring economic benefits
d. immigrants reduce labor shortages

7) What is Simon's attitude toward immigration?

a. restrict it
b. keep it unrestricted
c. restrict immigration of unskilled workers
d. set quotas based on the needs of industry

8) Which of the following statements does Graham challenge?

a. people who want to limit immigration are racists
b. the country is overpopulated
c. the resources of the nation are dwindling
d. immigration has become a central question in American politics

9) What was the subject of the magazine article titled, "The Browning of America"?

a. immigration
b. water shortages
c. air pollution
d. power failures

10) What does Simon refer to as a "one-time benefit?"

a. immigrants arrive without elderly parents who need social security
b. immigrants do work no one else will do
c. immigrants bring desperately needed capital into the country
d. immigrants have skills the country needs

11) Which statement best expresses Graham's view on the importance of history?

a. history teaches by analogy
b. history repeats itself
c. history is the best guide for policy
d. history may be misleading

12) What did President Nixon's 1972 commission conclude?

a. the nation's resources could support open immigration
b. current standards of life are in jeopardy
c. the U. S. population will stabilize if there is no further immigration
d. most immigration opponents are racists

Name _____

Attitude Measure
(Reading Comprehension)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Name _____

WRITING ASSIGNMENT

Imagine that it is 1981 and you are an educated citizen living in California. You are very interested in immigration and have just read the testimonies about immigration by Julian Simon, a professor of sociology, and Otis Graham, Jr. a professor of history. A friend asks you about some of the concerns about immigration in the country.

Write an essay in which you explain the most important ideas and issues your friend should understand. Your essay should be based on two major sources: (1) the general concepts and specific facts you know about American History, and especially what you know about immigration to this country; and (2) what you have learned from the readings.

Be sure to show the relationships among your ideas and facts.

Name _____

Attitude Measure
(Essay)

For each statement below, circle the number that shows how you felt when you were completing the prior knowledge measure.

Use This Scale:

1 = I did not feel this way
2 = I felt this way only a little
3 = I felt somewhat this way
4 = I felt strongly this way
5 = I felt very strongly this way

Circle one:

1.	1	2	3	4	5	I felt regretful.
2.	1	2	3	4	5	I was afraid that I should have been better prepared for this task.
3.	1	2	3	4	5	I felt that others would be disappointed in me.
4.	1	2	3	4	5	I felt I did not do as well on this task as I could have done.
5.	1	2	3	4	5	I did not feel very confident about my performance on this task.

Thinking Questionnaire

Name: _____ Sex: _____

Date: _____

Directions: A number of statements which people have used to describe themselves are given below. Read each statement and indicate how you thought while writing the essay. Find the word or phrase which best describes how you thought and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which seems to describe how you thought during the task.

		Not at All	Somewhat	Moderately So	Very Much So
1.	It was important to do well on the task.	1	2	3	4
2.	I asked myself questions about the task.	1	2	3	4
3.	I reworded the task so I could understand it better.	1	2	3	4
4.	I was aware of my ongoing thinking processes.	1	2	3	4
5.	I tried to do my best on the task.	1	2	3	4
6.	I carefully planned my course of action.	1	2	3	4
7.	It was ok to guess on the task since it did not count.	1	2	3	4
8.	I had multiple thinking techniques or strategies to solve the task.	1	2	3	4
9.	I was aware of my own thinking.	1	2	3	4
10.	I thought through the meaning of the essay assignment before I began to write.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr. General Immigration

	Not at All	Somewhat	Moderately So	Very Much So
11. I asked myself questions about what a task required me to do before I did it.	1	2	3	4
12. I judged the correctness of my work.	1	2	3	4
13. I evaluated the appropriateness of the thinking techniques or strategies that I used.	1	2	3	4
14. I was aware of which thinking technique or strategy to use and when to use it.	1	2	3	4
15. The task was challenging and interesting even if it didn't count for a school grade.	1	2	3	4
16. I tried to determine what the task was and what it required.	1	2	3	4
17. I knew how to recover from errors.	1	2	3	4
18. I selected and organized relevant information to solve the task.	1	2	3	4
19. I was aware of my trying to understand the task before I attempted to solve it.	1	2	3	4
20. I concentrated fully when doing the task.	1	2	3	4
21. I determined how to solve the task.	1	2	3	4
22. As I proceeded through the task, I asked myself, how well was I doing?	1	2	3	4
23. I kept track of my progress and, if necessary, I changed my techniques or strategies.	1	2	3	4
24. When I solved the task I was aware of checking how well I was doing.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

	Not at All	Somewhat	Moderately So	Very Much So
25. I worked hard on the task even if it did not count.	1	2	3	4
26. I set useful goals for myself.	1	2	3	4
27. I kept track of my progress.	1	2	3	4
28. The task was useful to check my knowledge.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr. General Immigration

Self-Analysis Questionnaire

Directions: The next sentences are about your feelings or thoughts when you were taking the task. Read each statement and decide how you felt during the task. Find the word or phrase which best describes how you felt and mark your answer sheet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which best describes how you felt during the task.

	Not at All	Somewhat	Moderately So	Very Much So
29. I felt regretful.	1	2	3	4
30. I was afraid that I should have studied more for this test.	1	2	3	4
31. I felt that others would be disappointed in me.	1	2	3	4
32. I felt I may not have done as well on this test as I could.	1	2	3	4
33. I did not feel very confident about my performance on this test.	1	2	3	4

Copyright © 1991 by Harold F. O'Neil, Jr.

Name_____
School_____
Teacher_____

STUDENT QUESTIONNAIRE

This two day testing period tried out some new ways to see what students know about history. Please give us your honest feelings on the questions below. We will keep your answers confidential and your teacher will not see them.

Circle the letter of the best answer for each question below.

1. Which word best describes the five periods you have spent on this project? Choose one.

   a. interesting
   b. tiring
   c. confusing
   d. boring

2. Which part of the testing did you like best?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

3. Which part did you like least?

   a. the first test: 20 short answer items
   b. the second test: multiple choice test on the reading passages
   c. semantic mapping
   d. the last test: the essay
   e. the whole thing

4. Overall, how was your performance on the tasks over these two days?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

5. On the first test with 20 short answer items, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

6. On the multiple choice test about the reading passages, how was your performance?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

7. How was your performance on semantic mapping?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

8. How was your performance on the essay?

   a. very good
   b. pretty good
   c. okay
   d. not very good
   e. pretty bad

9. Which of these statements best describes how you wrote the essay?

   a. I was able to put together new ideas from the reading with other facts and ideas I already knew.
   b. I focused mostly on writing style
   c. I used yesterday's reading passages for most of my ideas
   d. I mostly used information I already knew before I read the passages.

10. Where did you learn most of the information you used on the 20-item short answer test?

    a. the teacher's lectures
    b. the textbook for this class
    c. other classes I've had
    d. televisions
    e. other reading outside of school
    f. I don't know

11. How well do you think these tests show what you know about this subject?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

12. How hard did you try on these tasks?

    a. very well
    b. pretty well
    c. not very well
    d. not well at all

13. Compared to how you usually try on tests your teacher gives, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

14. Compared to standardized tests you take once a year or so, how hard did you try on these tasks?

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

15. Compared to homework,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

16. Compared to regular class discussions,

    a. much harder on these tests
    b. a little harder on these tests
    c. about the same on these tests
    d. a little less hard on these tests
    e. much less hard on these tests

17. The essay directions asked you to imagine you were in the same historical period as the authors of the reading passages and to explain the meaning of events to another person. How did that part of the directions influence your performance on the essay?

    a. it helped me organize and choose information in my writing
    b. it neither helped or hurt
    c. it interfered with my writing
    d. it made it fun
    e. I didn't pay any attention to it at all

18. Which of these statements is true for you? Check all that apply

    a. I didn't know the information the test asked for
    b. The essay asked me to write in a way that I haven't been taught to do
    c. These tests were more fun than regular tests
    d. These tests were about the right level of difficulty

19. No matter how I did, I feel like I know the topic tested

    a. very well
    b. pretty well
    c. only a little
    d. not at all

20. How do you rate yourself as a writer?

    a. I am a very good writer
    b. I am a pretty good writer
    c. I am a fair writer
    d. I am not at all good at writing

21. How good do you think you are in history?

    a. I am very good in history
    b. I am pretty good in history
    c. I am fair in history
    d. I am not at all good in history

22. How good a student are you?

    a. I am a very good student
    b. I am a pretty good student
    c. I am a fair student
    d. I am not a good student

23. How do you think you could improve your performance on the essay task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

24. How do you think you could improve your performance on the semantic mapping task?

a. much more time
b. much more instruction on how to do it
c. more practice
d. I could try harder
e. given opportunity to review some history content I have just learned
f. no way I can improve
g. some other way (please specify) _____

_____

25. Which sequence did you follow?

a. read first two texts, made semantic map, then read additional texts
b. read first two texts and additional texts, then made semantic map
c. read first two texts, then made semantic map while exploring other texts
d. other (please specify) _____

_____

26. Circle the appropriate number

How often do you:	At least once a week	A few times a week	once a month	less than once a month
A. Take tests in history readings	1	2	3	4
B. Write in-class essays in history	1	2	3	4
C. Have homework in history	1	2	3	4
D. Write in-class essays in other classes	1	2	3	4
E. Write history essays at home	1	2	3	4
F. Do longer research papers in history	1	2	3	4
G. Take short answer tests in history	1	2	3	4
H. Take multiple-choice tests in history	1	2	3	4

Appendix D

Essay Scoring Rubric

Essay Scoring Guidelines

1. **General Content Quality (GCQ)**

   How well does the student know and understand this historical content? (0 - 5 point global rating: 0 = no response, 5 = highest level of understanding)

2. **Prior Knowledge: Facts and Events (PK)**

   This is a measure of the extent to which the student incorporates relevant concrete information that is not mentioned in the speeches into his or her essay. This type of information may include pieces of legislation, court decisions, names of people, places or events, and general information about the period.

   Statements of opinion are not included (e.g., "Lincoln was our greatest president"). The student should not be penalized for information that is incorrect (e.g., "In 1770 some people moved from England to America and started the first settlement there"). (Mistakes will be accounted for in another scale.) Extremely common knowledge such as "slaves came from Africa" is not counted in this context.

   Score point guidelines:

   0 – no response
   1 – no facts/events mentioned that are not found in the texts of the speeches
   2 – one to two facts/events
   3 – three to four facts/events
   4 – five to six facts/events
   5 – seven or more facts/events

   Example: At Harper's Ferry John Brown attempted to lead a slave revolt but failed.

3. **Principles/Concepts - Number (PN)**

   This is a measure of the number of different social studies concepts or principles that the student uses with comprehension.

   A concept is an abstract, general notion, such as "inflation." It does not refer to particular events or objects (such as one particular period of inflation) but instead represents features common to a category of events or objects. "Imperialism," for example, does not refer to any specific facts or events; it is a heading that characterizes a class of behaviors and beliefs. "Industrialization" likewise identifies a class of activities and events that share common properties. It must be clear that the student is using a term conceptually, not just as a label.

A principle is a rule or belief used to justify an action or judgment, as in the statement "Slavery is immoral," where "morality" serves as a justifying principle.

It should be evident that the student understands the principle on a conceptual level. Concepts or principles should not simply be mentioned within a quotation or paraphrase from the text with no clear indication of understanding.

To earn a score point, the concept or principle need not be named explicitly, as in "Constitutionality was an important principle that influenced the debate over slavery."

Score point guidelines:

0 – no response
1 – no principles/concepts
2 – one principle/concept
3 – two principles/concepts
4 – three principles/concepts
5 – four or more principles/concepts

4. **Proportion of Text Detail (TX)**

This is a measure of the amount of material from the text of the speeches that is used in the essay.

A text detail is a quotation, paraphrase, or any other reference to information and ideas in tests provided. It should be clear that the text detail in the student essay was extracted or learned from the texts provided. If you believe that the student did not obtain information from the texts provided, do not count it as a text detail.

Assign the score point which comes closest to the proportion of text detail in the student's essay. (E.g., In a long essay, 1 or 2 sentences reflecting information from the provided text will earn a TX score of 1.)

Score point guidelines:

0 – no response
1 – no information from the text
2 – material from the text accounts for about 1/4 of the essay
3 – material from the text accounts for about 1/2 of the essay
4 – material from the text accounts for about 3/4 of the essay
5 – the essay uses or is based on material from the text only

5. Misconceptions (MI)

This is a measure of the amount of incorrect information, or the number of misconceptions or misinterpretations, in the essay. A higher score indicates more errors.

There are three possible types of errors to consider:

– factual errors such as incorrect names or dates
– misconceptions about the historical period
– misunderstanding or misinterpretation of the text of the debates

These errors should be scored accordingly:

0 – no response
1 – no misconceptions
2 – very minor error or misconception
3 – several minor errors and/or a moderate misconception
4 – at least one serious misconception
5 – one or more serious misconceptions central to the essay

Wrong opinions or judgments (e.g., "Douglas made good arguments and Lincoln didn't know what he was talking about") are not counted as misconceptions.

6. Argumentation (A)

How well does the student organize historical knowledge to develop a convincing argument or interpretation?

(0 - 5 point global rating: 0 = no response, 5 = highest level of coherent and cohesive argumentation)

This scale focuses on how well the student analyzes and organizes historical evidence to make a well-reasoned argument or interpretation. Essays scoring at the highest level will provide adequate evidence to support and justify interpretive stances and a chain of logical argumentation or analysis. In the development of a plausible analysis or argument, coherence will naturally be important. A paper that lacks coherence or logical flow should not earn the highest score.

Appendix E

Summary of Analysis of Variance Repeated Measures on Topics and Raters

Table A

Summary of Analysis of Variance Repeated Measures on Topics and Raters
General Content Quality (GCQ)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	4655.67	1	4655.67			5.70
Subjects (S)	250.31	67	3.74			0.22
Topics (T)	5.95	2	2.98			0.01
Raters (R)	3.88	3	1.29			0.00
ST	124.44	134	0.93	5.02	0.00	0.19
SR	59.25	201	0.29	1.59	0.00	0.04
TR	6.03	6	1.01	5.43	0.00	0.01
STR	74.42	402	0.19			0.19

Table B

Summary of Analysis of Variance Repeated Measures on Topics and Raters
Prior Knowledge–Facts and Events (PK)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	2586.30	1	2586.30			3.16
Subjects (S)	117.36	67	1.75			0.09
Topics (T)	2.11	2	1.06			-0.00
Raters (R)	30.68	3	10.23			0.05
ST	78.23	134	0.58	2.67	0.00	0.09
SR	60.05	201	0.30	1.36	0.01	0.03
TR	4.19	6	0.70	3.19	0.01	0.01
STR	88.03	402	0.22			0.22

Table C

Summary of Analysis of Variance Repeated Measures on Topics and Raters
Principles/Concepts–Number (PN)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	1714.42	1	1714.42			2.08
Subjects (S)	126.21	67	1.88			0.07
Topics (T)	24.51	2	12.26			0.04
Raters (R)	7.35	3	2.45			0.01
ST	129.38	134	0.97	6.61	0.00	0.20
SR	50.15	201	0.25	1.71	0.00	0.03
TR	6.93	6	1.16	7.91	0.00	0.01
STR	58.74	402	0.15			0.15

Table D

Summary of Analysis of Variance Repeated Measures on Topics and Raters
Proportion of Text Detail (TX)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	9293.33	1	9293.33			11.34
Subjects (S)	392.39	67	5.86			0.28
Topics (T)	45.32	2	22.66			0.06
Raters (R)	64.73	3	21.58			0.08
ST	323.91	134	2.42	5.77	0.00	0.50
SR	101.66	201	0.51	1.21	0.06	0.03
TR	28.72	6	4.79	11.43	0.00	0.06
STR	168.31	402	0.42			0.42

Table E

Summary of Analysis of Variance Repeated Measures on Topics and Raters
Misconceptions (MI)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	2526.52	1	2526.52			3.06
Subjects (S)	142.18	67	2.12			0.03
Topics (T)	50.12	2	25.06			0.07
Raters (R)	28.01	3	9.34			0.02
ST	219.38	134	1.64	2.92	0.00	0.27
SR	145.86	201	0.73	1.29	0.02	0.05
TR	25.34	6	4.22	7.53	0.00	0.05
STR	225.57	402	0.56			0.56

Table F

Summary of Analysis of Variance Repeated Measures on Topics and Raters
Argumentation (A)

Source	Sum of Squares	*DF*	Mean Square	*F*	Probability	Estimate of Variance
Mean	4411.13	14411.13			5.39	
Subjects (S)	213.58	67	3.19			0.19
Topics (T)	15.42	27.71			0.02	
Raters (R)	16.65	35.55			0.02	
ST	112.23	134	0.84	4.33	0.00	0.16
SR	60.73	201	0.30	1.56	0.00	0.04
TR	14.05	62.34	12.11	0.00	0.03	
STR	77.75	402	0.19			0.19